

**Confronting Discrimination and
Unravelling the Veil of Prejudice**

Martin Blaakman

θπ



Quaestiones Infinitae

Publications of the Department of Philosophy and Religious Studies

Utrecht University

VOLUME 134

© 2021 Martin Blaakman

ISBN 978-94-6103-090-0

Printed by ProefschriftMaken

Cover design by Sara Terwisscha van Scheltinga, persoonlijkproefschrift.nl

Confronting Discrimination and Unravelling the Veil of Prejudice

The epistemic conditions of responsibility for hidden
prejudices

Discriminatie Bestrijden en de Sluier van Vooroordelen Ontrafelen
De epistemische voorwaarden van verantwoordelijkheid voor
verborgen vooroordelen
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 10 september 2021 des middags te 12.15 uur

door

Martin Jacobus Blaakman

geboren op 27 juni 1973
te Oostburg

Promotoren:

Prof. dr. H.H.A. van den Brink

Dr. J.H. Anderson

Beoordelingscommissie:

Prof. dr. I.A.M. Robeyns

Prof. dr. M.M.S.K. Sie

Prof. dr. J.A.M. Bransen

Dr. A. Kalis

Dr. P. Robichaud

Contents

Acknowledgements	7
1. General introduction	9
1.1. Confronting practices of hidden prejudice	11
1.2. The research question	16
1.3. Methodological approach and research aims	27
1.4. Two frameworks: mentalism and pragmatism	31
1.5. Context and relevance	38
1.6. Outline of chapters	44
2. The liberal notion of responsible agency	47
2.1. Introduction	49
2.2. Reflective volitional autonomy	49
2.3. The problem of manipulation	59
2.4. Reflective epistemic autonomy	71
2.5. The mentalist roots of reflective epistemic autonomy	80
2.6. Conclusion	87
3. The veil of prejudice	89
3.1. Introduction	91
3.2. Implicit bias: a mentalist notion of hidden prejudice	93
3.4. Prejudice as ignorance	98
3.5. Prejudice as ignorant resistance	102
3.6. Ignorant resistance as a social practice of censorship	105
3.7. Social practices of silencing	108
3.8. Conclusion	113
4. The mentalist framework: the impossibility of responsible agency	115
4.1. Introduction	117
4.2. The determinism thesis	119
4.3. The manipulation argument	126
4.4. The reflective belief challenge	133
4.5. The manipulation argument revisited	140
4.6. Conclusion	151
5. A pragmatist framework: freedom and rationality	153
5.1. Introduction	155
5.2. Free will as the wrong notion of freedom	156

5.3.	Transcendentally free but non-responsible actions	166
5.4.	Brandom's radical position	171
5.5.	Rationality as normative bindingness	176
5.6.	Normative bindingness as practical	185
5.7.	The Cognitive Suicide Argument	195
5.8.	Conclusion	205
6.	A pragmatist notion of the veil of prejudice	207
6.1.	Introduction	209
6.2.	The space of reasons as a social space	210
6.3.	Bourdieu's 'learned ignorance'	219
6.4.	Hidden prejudice as implicit intentionality	228
6.5.	Implicit intentionality as social intentionality	239
6.6.	Conclusion	246
7.	A pragmatist notion of responsible agency	249
7.1.	Introduction	251
7.2.	Epistemic autonomy revisited	251
7.3.	Freedom revisited	260
7.4.	Epistemic responsibility for social practices of silencing	267
7.5.	Non-deliberative action as non-manipulative influencing	273
7.6.	Non-deliberative action as narrative performance	279
7.7.	Narrative performance as confronting discrimination	287
7.8.	Responsibility for hidden prejudices: criticising the mentalist approach	300
7.9.	Conclusion	314
8.	Conclusion	317
8.1.	One research question...	319
8.2.	...and two answers	320
8.3.	Evaluating both frameworks	323
8.4.	Suggestions for future research	328
	Appendices	331
	Bibliography	333
	Summary	389
	Samenvatting in het Nederlands	398
	Curriculum Vitae	409
	Quaestiones Infinitae	410

Acknowledgements

This dissertation is the harbour where I arrived after a long journey. From a personal perspective, this journey feels to me like a lifetime connecting birth and death. During this period, I became the father of a son, and I lost my father-in-law; I became the father of a second son, and I lost my own father. And while writing has primarily been a solitary process, I have not travelled without companions. In fact, I had and still have many companions whom I would like to thank.

First, I want to thank Veit Bader. He was my first supervisor. He gave me the chance to become an external PhD candidate. Without his trust in me, patience, and encouragement, I could not have started or continued writing my thesis. Also, I owe many thanks to Bert van den Brink, who was so kind to take over Veit's supervision, always supporting me. Later, Joel Anderson joined Bert in supervising me. To him, too, I owe many thanks. He gave me good thoughtful comments. Without Bert and Joel, I could not have continued and finished writing my thesis.

During most of these years, I worked for the National ombudsman. In my spare time, I worked on my thesis. I want to thank the National ombudsman, now my former employer. He consented to sabbaticals that allowed me to make progress on my thesis.

Writing a dissertation can be very relaxing when you take a break, once in a while: meeting with friends, going to the cinema, discussing, eating, and drinking with them. I want to thank them all. I want to express special gratitude to my paranymphs, Nikki Smaniotto and Roland van Loosbroek. During these years, Roland has often helped me with my thesis: reading and commenting on chapters, discussing philosophical ideas and arguments. Nikki has always believed in me. Also, thanks to Geor Hintzen, who read and commented on parts of my thesis. I also thank my other friends who know me for so long, particularly: Sandra, Alvio, Klaas-Jan, Hanneke, Nina and Rita. They have supported me throughout the years.

My family, too, stood by my side. Special thanks to Ingrid, my mother-in-law, who was always helpful to Marjolein and me. I would also like to thank my brothers- and sisters-in-law. And special gratitude to my brothers, Angelo and Steven.

To my parents, Ria and Marcel, I owe more than I can express. When I was 17 years old and chose to study philosophy, they did not know what philosophy meant. Nevertheless, they have supported me, always, and always. I had wished that my father could have witnessed the completion of my dissertation. His perseverance was an inspiration to me. *Mama, je doet het goed. Jij en je drie zoons gaan verder.*

Marjolein, my loved one: thank you so very much. You are my love and my companion in life. We are the parents of two great boys: Jonas and Simeon, whom I am so proud of. It has taken a long time to complete my dissertation. You have been patient with me. All these years. More than once, you took care of our boys while I made another attempt to finish it. Now, it is time to close the book. Let us go out and watch our boys play.

1

Chapter 1

General introduction

1. General introduction

‘It isn’t necessarily through acts of physical violence – lynching, mob attacks, or slaps in the face, whether experience first-hand or by word of mouth – that a child is initiated into the contradictions of segregated democracy. Rather, it is through brief impersonal encounters, stares, vocal inflections, hostile laughter, or public reversals of private expectations that occur at the age when children are most receptive to the world and all its wonders.’

Ralph Ellison (1989:821)

1.1. Confronting practices of hidden prejudice

Non-discrimination is an essential value to modern liberal democracies. It holds the promise for equal treatment of their citizens and has therefore motivated both citizens and politicians in confronting discrimination. However, confronting discrimination provides us with a dual challenge of understanding: understanding how it works to confront it more effectively and understanding what actions of confronting discrimination may count as morally justified.

In the next chapters, I will develop an argument along these two lines by focussing on two themes: *hidden prejudices* and *non-deliberative actions*. In this section, I will only provisionally introduce them by presenting two examples of confronting discrimination: the first example will illustrate the phenomenon of hidden prejudices, the second the phenomenon of non-deliberative action. Only in the next section will I look more closely at both phenomena and introduce my research question.

The case of DH and others v. The Czech Republic

Between 1996 and 1999, some 18 children living in Ostrava, one of the largest cities in the Czech Republic, had been placed in special schools for children who have mental deficiencies and cannot attend regular primary schools. The oldest children had been somewhere between the ages of 11 and 14, but most of them were between the ages of 5 and 8. Usually, in that period, when a child entered the school for the first time or if the school observed difficulties in its ordinary primary-school education, it was tested for its intellectual capacities in an educational psychology centre. Furthermore, whenever the test results indicated a mental disability, the centre could propose to the head of the school that the child was placed in another school, a school for special education. Only if the centre had provided the head of the school with the relevant documents and only if the parents consented to this proposal could the head of the school make the final decision for placement. These standard rules had been followed for the 18 Ostrava

children.¹ At the same time, the rules had been neutral and involved no reference to grounds that might be considered discriminatory. Nevertheless, in 1999 on behalf of these children, a constitutional appeal was lodged with the Czech Constitutional Court against the Czech Republic, complaining that they had been discriminated against in the general functioning of the special-education system. Later that year, the Constitutional Court dismissed their appeal.

Who were these children? They belong to a group that nowadays has a national-minority status in the Czech Republic: the Roma. But the residence of Roma is not restricted to the Czech Republic. Roma people are the largest minority of the European Union (EU), with an estimated 6 million people.² They live all over the European Union, but mainly in the Eastern European member states. Even since various Eastern European countries wanted to become, and finally became, member states of the European Union, many institutions in the European Union have begun to recognise the urgency to confront discrimination against Roma people as they live in many places in the European Union as second-class citizens. It inspired these institutions – amongst others the Open Society Institute and the Council of Europe – to effectively start reducing practices of prejudices in what is sometimes called the ‘Decade of Roma inclusion’ (2005-2015). It also motivated the European Union member states to adopt an ‘EU Framework for National Roma Integration Strategies up to 2020’ (European Union 2011).

However, despite such efforts, Roma people seem to remain invisible to their fellow Europeans as a people who ‘are victims of prejudice and social exclusion, despite the fact that EU countries have banned discrimination’, as a website of the EU states.³ The invisibility of such practices of prejudices is illustrated by the legal procedure on behalf of the 18 Ostrava children. Theirs is the *Case of DH and others v. The Czech Republic*, also known as the *Ostrava case*, after the Czech city which provided the stage for this case. After their appeal was dismissed by the Constitutional Court, their case was brought before the European Court of Human Rights. To be more specific: before the Chamber (not to be confused with the Grand Chamber) of the Second Section of this court. So how was the claim of discrimination supported? While the rules for placement in schools for special education seemed neutral and the psychological tests seemed

1 These procedures are described in Decree no. 127/1997. See paragraphs 34–36 of the decision by European Court of Human Rights (2007), *Case of DH and Others v. the Czech Republic*, 13 November 2007, application number 57325/00.

2 By comparison, some 12 countries within the EU have a smaller population.

3 See https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/roma-and-eu/roma-integration-eu_en (accessed on 3 September 2020). The term ‘Roma’ might suggest a (for example, ethnically, culturally) homogenous people. This is not the case. See the remark by the European Commission (2012:2, note 1): ‘The term “Roma” is used here, as well as by a number of international organisations and representatives of Roma groups in Europe, to refer to a number of different groups (such as Roma, Sinti, Kale, Gypsies, Romanichels, Boyash, Ashkali, Egyptians, Yenish, Dom, Lom) and also includes Travellers, without denying the specificities and varieties of lifestyles and situations of these groups’.

objective, it was argued on behalf of the 18 Ostrava children that this was not the case. Decisions to place Roma children in schools for special education were ‘not based on any actual mental disability but rather on language and cultural differences which were not taken into account in the testing process’⁴ because the psychological tests ‘had been devised solely for Czech children and had not recently been standardised or approved for use with Roma children’.⁵ At the same time, ‘no measures had been taken to enable Roma children to overcome their cultural and linguistic disadvantages in the tests’.⁶ And therefore, decisions for placement, based on such tests, ‘resulted in *de facto* racial segregation and discrimination that were reflected in the existence of two separately organised educational systems for members of different racial groups’.⁷ To support their claims, the European Roma Rights Centre, which represented the 18 Roma children, had presented statistical evidence showing that ‘a Roma child in Ostrava was 27 times more likely to be placed in a special school than a non-Roma child’.⁸ However, the European Court, just like the Constitutional Court before, did not recognise that the practice of placing Roma children in a special school was a discriminatory one. In its decision of 7 February 2006, the Chamber of the Second Section dismissed the statistical evidence, considering that, consistent with earlier jurisprudence, ‘statistics are not by themselves sufficient to disclose a practice which could be classified as discriminatory’.⁹

However, the case was still not closed for the 18 Ostrava children. On behalf of them, it was brought again before the European Court, this time before its Grand Chamber, the last option for appeal. Clearly, the judges had had their discussions and disagreements about whether statistical evidence would be enough, as can be seen in the dissenting opinions of four judges. Nevertheless, on 13 November 2007, the Grand Chamber of the European Court of Human Rights decided that the claim of the 18 Ostrava was correct.¹⁰ This decision was welcomed as a significant and much-needed success for the Roma people.

4 See European Court of Human Rights, *Case of DH and Others v. the Czech Republic*, judgement of 13 November 2007, paragraph 167.

5 See European Court of Human Rights, *Case of DH and Others v. the Czech Republic*, judgement of 13 November 2007, paragraph 40.

6 See European Court of Human Rights, *Case of DH and Others v. the Czech Republic*, judgement of 13 November 2007, paragraph 40.

7 See European Court of Human Rights, *Case of DH and Others v. the Czech Republic*, judgement of 13 November 2007, paragraph 25.

8 Here I quote from the decision (see note 6) of 13 November 2007 (see note 6), paragraph 18, but the evidence was presented already, but not mentioned explicitly, in the decision of 7 February 2006. It should be noted that the overrepresentation of Roma children in schools for special education is not unique for the Czech Republic, but also occurs in primarily eastern European countries: see, e.g., European Union Agency for Fundamental Rights (2014). The problem may, however, also obtain for western European member states of the European Union, for example, the Netherlands: see Hopman (2016) and Seidler et al. (2020).

9 Paragraph 46 of the Chamber of the Second Section of the European Court of Human Rights, *Case of DH and Others v. the Czech Republic*, Application no. 57325/00, judgement of 7 February 2006, available at <https://hudoc.echr.coe.int>. For a critical discussion of this decision: see Goodwin (2006).

10 See previous note for this decision.

The Ostrava case provides, I would like to suggest, an example of *hidden prejudices*: prejudices that are not manifest in what people say, not evident in rules, procedures or documents, but are somehow, somewhere encrypted in people's psychology, inducing them to perform actions that by themselves seem unprejudiced, but whose aggregated effects turn out to disadvantage people from an already disadvantaged minority. The phenomenon of hidden prejudices might explain the difficulties various judges from various courts had in assessing whether the practices in the Ostrava case were discriminatory.

The case of Brown v. Board of Education

On a different continent, in a different decade, a different court had taken a decision in the case of *Brown v. Board of Education*. As the American Civil War had ended, the equality of all citizens before the law was then finally also recognised for African Americans. Or so it seemed, because ever since the laws and social practices on race became dominated by the 'separate and equal' doctrine: it stated that the separation of black and white people was compatible with the idea of equality. The legal source for this doctrine was a decision of the Supreme Court of the United States of America in the case of *Plessy v. Ferguson* (1896).¹¹ However, in 1954, it was repudiated by the Supreme Court in the landmark case *Brown v. Board of Education*, arguing that state laws establishing separate educational facilities are unconstitutional.¹² The decision was celebrated as a major victory.

However, this decision was not enough for the effective desegregation of education. Practices of prejudices against African Americans remained as strong as ever. These were not hidden, as in the Ostrava case, but clear and loud in what white people publicly said and did. This is illustrated in one powerful image that in those years was published in the press. It shows a young black woman, dressed in a neat white dress with a checkered pattern at the bottom, wearing dark glasses, walking straight on in an almost stoic fashion, seemingly untouched by the 'hundreds of Little Rock citizens' (as it was reported earlier that day on the radio) that surrounded her, primarily women, walking along with her, coming up to her but not touching her, and yet clearly performing some act of non-physical violence by looking at her with grim faces, while one woman in particular, firmly walking some steps behind her as if pursuing her, is shouting at her with a speechless mouth.¹³

The famous photograph of this scene shows Elizabeth Eckford, a fifteen-year-old

11 *Plessy v. Ferguson*, 163 US 537 (1896).

12 *Brown v. Board of Education of Topeka, Kansas*, 347 US 483 (1954). What is usually referred to as *Brown v. Board of Education* includes not just this case but also the following cases: *Bolling v. Sharpe* (1954) and *Brown v. Board of Education* (1955). For a detailed historical analysis of these cases: see Kluger (1975).

13 For an inside description of what happened that day: Beals (1994).

student who, on 4 September 1957, the first day of her new school.¹⁴ Although the case *Brown v. Board of Education* recognised for African Americans the right to equal education, it took a few years before the first African American students dared to oppose the practices of prejudice and exercise their newly acquired right by registering for schools that were now formally open to them but still had a population of only white students. One of them was Elizabeth Eckford. She had wanted to attend a better school. However, better schools were usually, not surprisingly in those days, white schools, one of which was Little Rock High School, a public school in Arkansas, United States. As the photograph shows, entering the school building on her first day of school, all on her own, she met with fierce opposition from white people. The powerful image of injustice committed against her right to equal education aroused public indignation across the nation.

The first African American students that visited ‘white’ schools belonged to the broader group of African Americans that faced their situation as second-class citizens but nevertheless struggled for their rightful place within society. Their struggles were part of what became known as the Civil Rights Movement. Those involved in this movement and in earlier struggles for equal treatment had initially used the public and legal means that were supposedly available to all citizens when they wanted to voice and denounce injustices. Crucial to both the effectivity and the legitimacy of these means is, of course, *deliberation*, as the ability to express one’s claims on injustices and to support them with reasons, but also as an essential value to liberalism as a political philosophy. African Americans exercised their deliberative abilities in using these means: within the public debate and, more specifically, within the courts. However, as they continued to perform their deliberative actions, they experienced how, even in a democracy, deliberation may come to an end and lose its effectiveness and legitimacy. They, therefore, started looking for other means to address injustice and finally resorted to ‘non-violent direct actions’.¹⁵

About a decade later, Rawls dedicated a few sections of his *A Theory of Justice* (1971) to the phenomenon of non-violent direct actions: civil disobedience, as Rawls preferred to call them. He recognised that it might be justified if a minority finds itself in a situation in which, as Rawls describes it, ‘the normal appeals to the political majority have already been made in good faith and they have failed. The legal means of redress have proved of no avail’ (1971:373). Rawls was sensitive to the special, *moral* questions that ‘non-violent direct actions’ raised and which specifically consisted in

14 Of the attempts by Elizabeth Eckford to enter her new school, several photos were taken. The most well-known was taken by Will Counts: see Counts (1999), whose cover shows that specific photograph. Another, lesser-known photograph, capturing the same scene at almost the same moment, but from a different angle, was taken by Pete Harris: see Berger (2013:20). Yet another photograph of Elizabeth Eckford but taken at another moment during her walk through the crowd, picturing different people around her, was the starting point for Arendt’s reflections on the Little Rock event: Arendt (1957), which includes a reprint of the photo. For Arendt’s response: see the discussion of case study 1 in chapter 7, section 7.

15 See King (1963).

their disobedience to the law. He made the claim that under certain circumstances, non-violent direct actions may be justifiable and even effective. However, he failed to explicitly recognise that his claim, if it is correct, indicates the failure of deliberation within a democracy and that, therefore, non-violent direct actions may be effective, not because of its deliberative aspects, but because of its *non-deliberative* aspects. While non-violent direct actions are not completely cut off from normal, deliberative actions, they nevertheless constitute their own category of actions: ‘non-deliberative actions’, as I want to call them. It is for these non-deliberative actions – for example, the illegal marches and sit-ins – that the Civil Rights Movement became known across the nation and across the world.

For the 18 Ostrava children, the victorious decision of 2007 came too late, much too late. By then, even the youngest children had at least reached the age of 16. However, it seems that also for children who came after them, not enough progress has been achieved.¹⁶ So how can Roma people with them move forwards? Some believe that the case *DH and others v. The Czech Republic* is the equivalent of the case *Brown v. Board of Education*.¹⁷ If this makes sense, then perhaps Roma people might also learn from African Americans what to do if legal victories prove to be inadequate to confront discrimination. So this is my suggestion: Roma might perhaps confront practices of hidden prejudice by performing non-deliberative actions. And then we return to the dual challenge of understanding I started with, a little more specified now: understanding how *hidden prejudices* work to confront their practices more effectively and understanding whether, in confronting discrimination, *non-deliberative actions* may count as morally justified.

In the following sections, I will introduce my research question (1.2.). Next, I will explain my methodological approach and research aims (1.3.), the two frameworks from which I will explore my research question (1.4.), and the context and relevance of my research question (1.5.). Finally, I will provide an outline of the chapters (1.6.).

1.2. The research question

Practices of prejudice have led African Americans and Romani people – and so many other minorities in many different societies – to become second-class citizens in modern liberal democracies. It raises enormous challenges for discriminated people – but also for politicians, policymakers and ordinary non-discriminated citizens – in finding out how such practices may be confronted. In the following chapters, I will be concerned with this issue as a *moral* one. I will limit my moral exploration, first, to practices of *hidden*

16 See, e.g. Amnesty International; European Roma Rights Centre; Open Society Fund and Czech Society for Inclusive Education (2015); (Ombudsman of the Czech Republic, officially known as) Public Defender of Rights of the Czech Republic (2019). Interesting information is also provided by a survey, set up and conducted by people from the Roma community itself: see Karvayová et al. (2016).

17 See, e.g. Goldston (2005); Goodwin (2009); Greenberg (2009-2010); Minow (2010).

prejudice and, second, to *non-deliberative* actions as a type of actions that might have the potential to reduce, in a morally justified way, those practices. In this section, I will explain its moral perspective and my choice for its two demarcations.

To explore the *moral* conditions of actions that confront practices of prejudice,¹⁸ I choose to discuss them in terms of responsibility.¹⁹ While the notion of responsibility is usually understood and analysed with an almost exclusive focus on ‘being responsible’, I will use this notion as having a dual structure, consisting of both ‘being responsible’ – or (the term that I prefer to use): ‘responsible agency’ – and ‘holding responsible’.²⁰ For the purpose of my research question, I will treat actions that (attempt to) reduce practices of prejudices as instances of holding responsible: such actions, I assume, are performed by those who view themselves as standing in a moral relation towards prejudiced people and who view their actions, in part at least, as justified in terms of that relation. It means that the moral conditions we are looking for are the conditions of what actions may count as holding responsible. I will assume that actions do not count as a form of holding responsible if they consist of *violence against persons*.²¹ So I will exclude these from my discussion.

The notion of responsibility still leaves the possibility of discussing the case of hidden prejudice from a normative but non-moral, for example, legal perspective. From such a perspective, it need not be interesting whether people upholding practices of hidden prejudice do or do not have a hidden prejudice. In American jurisprudence, there have been discussions for a long time about what is decisive for the assessment of discrimination: that the act had a *discriminatory impact* or that it originated in a

18 Here the verb ‘reduce’ is used, not in the sense that actions, whatever they are, will put an end to such practices, but in a more modest sense, that they will make those practices less frequent, mitigate the harshest practices among them and perhaps change some part of them into unprejudiced practices.

19 Various authors have observed the relatively recent arrival of the term ‘responsibility’ in our moral vocabulary, somewhere during the nineteenth century, as compared to other, older moral terms. See e.g. Niebuhr (1963: ‘late-born child’); Jonsen (1986: ‘new arrival’); Bayertz (1995a); W. Davis (2001); Parker et al. (2011). As far as I could find out, the first uses of responsibility as a ‘terminus technicus’ in any discipline can be traced back to the beginning of the nineteenth century, and even then, only in France, when it was introduced as a juridical notion in the ‘Code Civil’ of 1804 (for references on the history of ‘responsabilité civile’, see Lacroix 2009). As a distinctly philosophical notion, it appeared only for the first time, in English, French and German, in the second half of the nineteenth century (McKeon 1957). Weber seems to have been the first, in his lecture ‘Politik als Beruf’ (1919), to talk specifically about a *Verantwortungsethik*, an ethics of responsibility (De Villiers 2012).

20 See Macnamara (2011).

21 Such actions should be distinguished from ‘violence against objects’, i.e. the damage to property. For a careful explanation of this distinction: see Frankenberg and Rödel (1981, chapter IV). They should also be distinguished from coercion in the sense of the *threat* of force. See, e.g. Anderson (2010). Coercion in this sense still involves some minimum of negotiation. See, e.g. Bader (1991, chapter XI). To explore whether, for example, the damage to property or the threat of force is permissible as a form of holding responsible (as in practices of social protest) is beyond my present scope.

discriminatory purpose.²² What started them was the landmark decision of the Supreme Court *Griggs v. Duke Power* (1971).²³ In the jurisprudence of the European Union, the equivalent for this distinction is one between direct and indirect discrimination.²⁴ There may be good reasons to dismiss the issue of discriminatory purpose as irrelevant and instead assess whether the act had a discriminatory impact. The Ostrava case, for example, has been viewed as an example of indirect discrimination. As we have seen in the previous section, a court might use statistical evidence to decide such a case without being concerned about evidence for any discriminatory purposes.²⁵ I am interested, however, in a moral assessment of discriminatory practices and, therefore, in understanding whether or how practices that have a discriminatory impact may lack discriminatory purpose but may nevertheless be traced to hidden prejudices. It suggests that we might make a distinction, theoretically at least, between *practices of discrimination* and *practices of prejudices*: while the first category includes *all* practices with a discriminatory impact, regardless of their cognitive, psychological or social origins, the second includes only those discriminatory practices that *are rooted in prejudices*.²⁶ What will concern me here is only the second category.²⁷

Within this second category, I will focus further on only those discriminatory practices that might be rooted in *hidden* prejudices. But if a hidden prejudice is not a discriminatory purpose, then what is it? Here I suggest a preliminary concept of hidden prejudice. It will be outlined at a sufficiently abstract level to allow a discussion, in the next chapters, of more specific proposals. First, actions and practices may be understood, at a general level and in a simplified form, in terms of the following motivational pattern: some belief *A* leads to action *B*. Or simply: *A leads to B*. In cases of prejudiced actions, belief *A* presents a prejudice, and action *B* presents an act having a discriminatory impact on the person that is targeted by the prejudice. In standard cases of practices of prejudice, people *know* they have prejudiced beliefs, are motivated by them to perform

22 The *discriminatory impact* consists of unequal treatment based on religion, race, ethnicity, et cetera. If the intention of the agent had been precisely to realise this impact, this might be called the *discriminatory purpose*. The distinction between practices that only have a discriminatory impact and those that also have a discriminatory purpose is also known as a distinction between indirect and direct discrimination. See e.g. Lippert-Rasmussen (2015).

23 *Griggs v. Duke Power*, 401 US 431 (1971). For an extensive discussion of this decision: see Belton (2014). It introduced the view that what is decisive about assessing discrimination is the impact, also known as the *disparate impact theory*, ‘one of the most controversial developments in civil rights history’, as Belton puts it (2014:4).

24 This distinction, as Tobler (2008:5) remarks, ‘has been developed by the [European] Court of Justice through its case law since the 1960s, in order to enhance the effectiveness of EC non-discrimination law’.

25 See Tobler (2008); Devroye (2009); European Agency for Fundamental Rights (2018).

26 It is possible that someone has a prejudice, acts on this prejudice, but fails to realise a discriminatory impact. In that case, an ‘act of prejudice’ would not be an ‘act of discrimination’. I ignore, however, this possibility as theoretically uninteresting, at least for my argument.

27 To be clear about this point: whenever I talk about prejudices, I mean only those prejudices that have motivating force.

discriminatory actions and know they are motivated by them. Various psychological, political and sociological theories, however, have suggested the possibility that people may have a certain belief that motivates them to perform some action, but which they do not know of. Although various names have been used to express this phenomenon, here I will refer to it as one of people's *ignorance of their own actions as prejudiced*. This is how I will, for now at least, understand hidden prejudices.²⁸

When people are said to have hidden prejudices, it would mean that they do not know about *A* as a motivating belief for *B* or, more precisely, they are ignorant of the false grounds of some motivational belief *A* and even of the belief *A* itself. The idea of hidden prejudice as self-ignorance about one's own prejudices introduces the possibility of a boundary within people's self-knowledge: a boundary between the beliefs people know they have and the beliefs they do not know they have *about their own agency*; a boundary therefore between, as I will call it, *agential self-knowledge* and *agential self-ignorance*. It has been referred to by Charles W. Mills as a 'moral cognitive dysfunction' (1997:95).²⁹ People may therefore believe they are committed to ideals of equality, may even know that certain prejudices prevail in their society, and still perform discriminating practices as they are ignorant of the dispositions underlying their own practices. The issue of understanding hidden prejudice is, therefore, one of understanding the boundary between self-knowledge and self-ignorance.

The boundary that divides people's self-knowledge and self-ignorance – in whatever way we should qualify that boundary – will be called the *veil of prejudice*. This metaphorical image is closely related to two famous 'veil' metaphors that modern theory on justice has produced: one by Rawls, the best known, the other by Du Bois, lesser known, but probably more interesting for my purposes.³⁰ Rawls (1971) turned the veil into a metaphorical safeguard for justice when he argued that people, in a fictitious original position, should choose principles of justice 'behind a veil of ignorance'. For Rawls, the veil has an *epistemic* function because it 'excludes the knowledge of those contingencies [of their positions in real life] which sets men at odds and allows them to be guided by their prejudices' (1971:19). It helps people to know the principles of justice, to *truly see* justice, precisely because it functions as what *blinds*, and *should blind*

28 In chapter 3, section 2, I will provide some more historical background to this idea of ignorance and (hidden) prejudice.

29 In the context of social psychological research on hidden prejudice ('implicit bias') this has been called 'dissociation' (Washington and Kelly 2016) and 'divergence' (Johnson 2020). In recent philosophical discussions, but in a wider context, i.e. not specifically centred on practices of hidden prejudice, it has been called the 'belief-behavior mismatch' (Gendler 2008). A related notion is what Fricker (2007) has called a 'dissonance' between beliefs people may have and their perceptual judgements.

30 The veil as a moral-political metaphor was also used by Immanuel Kant (just once, as far as I can tell) in his *Die Metaphysik der Sitten* when he discusses the right of settlement for colonisers. In considering various arguments that might support this right, he rejects them because, as he remarks, one can easily see through this 'veil of injustice' (*Schleier der Ungerechtigkeit*), meaning that these arguments are intended to conceal, but in a too obvious way, their real motivating, immoral principle, namely that 'the goal justifies the means' (first part, second chapter, section 15; the edition I used: 1956:377).

, people from seeing their prejudices, analogous to how the blindfold functions for Lady Justice.³¹

For Du Bois, the veil is, in the first place, a metaphor for *exclusion*.³² He chose as the setting for his metaphor, not a fictitious situation, such as Rawls' original position, but the real lives of people, of black and white people in the racial divide of the United States. Himself of African American descent, he tells us about one of his childhood's memories, when he as a very young boy played with various children and how one of them, through her reactions to him – a refusal, a glance – made him realise, not just that *they* were white and *he* was black but even more so that *he* was 'shut out from their [i.e. white people's] world by a vast veil' (1903/ 2007:3). This is the discovery of what Du Bois elsewhere calls the 'Veil of Race' or 'Veil of Color'.³³

As I understand and interpret Du Bois' use of the veil, it also is an *epistemic* metaphor because the veil does not blind African Americans. They can see through it. They can see 'this American world' and how white people go about in this world. They can even see themselves and how white people treat them. But when African Americans see themselves, they look at themselves 'through the eyes of others, of measuring one's soul by the tape of a world that looks on in amused contempt and pity' (id.). Put differently: African Americans have consciousness of themselves but, living in a world dominated by white people, this consciousness is only mediated by the consciousness that white people have of them. This is what Du Bois calls, using yet another metaphor, 'double consciousness'. It suggests, next to the *outer* veil that divides African Americans from white people, a kind of *inner* veil – even if Du Bois himself does not use this particular phrase – one that divides African Americans from themselves, *against* themselves, because they are forced to live with two consciousnesses, 'two souls, two thoughts, two unreconciled strivings' (id.). Because the white world of which the African American is part, as Du Bois observes, 'yields him no true self-consciousness' (id.), this veil, therefore, excludes, for African Americans at least, knowledge of their 'true self-consciousness'. It hinders them from knowing about their experiences of injustice and *truly seeing* injustice in what white people *do* to them. Reflecting on his childhood memory, Du Bois admits that he 'had no desire to tear down that veil, to creep through' (id.), suggesting that this might

31 Rawls's use of the veil metaphor is consistent with the use of the veil (or blindfold) in traditional representations of justice. See, e.g. Resnik and Curtis (2011) although they suggest that the veil does not blind people: 'Both Rawls's veil and Justice's blindfold depend on the assumption that the wearer can, in fact, see but is committed to bounded knowledge' (Resnik and Curtis 2011:98). Interestingly Kant (see earlier note 24) uses the veil as a metaphor for *concealing* (i.e. *blinding* people to) injustice.

32 For a discussion on the origins of Du Bois's metaphorical use of the veil: see Schragger (1996).

33 The veil as a socio-political metaphor occurs at various places in Du Bois's writings. It is mentioned, for example, in the subtitle of his *Darkwater* [1920]: 'Voices from Within the Veil'. Probably the most important source for his use of 'veil' metaphors, including 'veil of race' and 'race of color', is his *The Souls of Black Folk* [1903].

be a possibility of solving the outer veil.³⁴ I think, however, that Du Bois' commitment to the idea of 'double consciousness' also commits him to the idea that simply tearing down the veil will not work, in part because the outer veil is also an inner veil. Even as African Americans come to 'see' the veil and to 'see' their own 'double consciousness', seeing does not yet present a solution to self-liberation.

In choosing the metaphor of veil of prejudice, I want to emphasise that the problem of hidden prejudices turns on understanding the *boundary* between agential self-knowledge and self-ignorance, a boundary for which the veil is only a provisional, metaphorical tool to understand what this boundary means. Rawls' and Du Bois' veil metaphors provide us with two different epistemic models for understanding this boundary. Rawls' veil of ignorance seems to play on a reversal of the values normally associated with seeing and being blind: while being blind is usually understood as a hindrance to achieving true knowledge, in the original position, it paves the way to knowledge of the principles of justice. But apart from this attempt at reversal, it shows that Rawls' metaphor draws on a basic, dichotomous, epistemic model: *knowing* is contrasted to *not knowing* just as *seeing* is opposed to a *radical not seeing* (namely *being blind*). For people with hidden prejudices, according to this epistemic model, the boundary between agential self-knowledge and self-ignorance about their prejudices is a blindfold: they are blind to their prejudices. Du Bois' use of the veil is more complex and more ambiguous than Rawls' use. Du Bois' veil of race, as I understand it, seems to lack the option of *not seeing* as a case of being blind. African Americans can see, but their seeing is a case of obstructed seeing, both of themselves and of the practices of prejudice performed by white people. And while seeing is obstructed and being blind is not an option, it is not obvious what *truly seeing* would consist in.³⁵ My interpretation of Du Bois' underlying epistemic model is one of *mastering* knowledge. To put it metaphorically: the veil functions as what allows people to see, even if at the same time it may obstruct their seeing. It is, therefore, our inevitable starting point in looking at our society and at who we ourselves are. Lifting the veil or tearing it down would therefore be no option because it would also take away the abilities to make sense of the world at all. Instead, we should try to solve our obstructed seeing, making it less obstructed, simply by unravelling bits of the veil and reweaving it. For people with hidden prejudices, according to this epistemic model, the boundary between self-knowledge and self-ignorance about their prejudices is a veil of seeing *and* obstructed seeing.

34 Du Bois uses the veil to describe his own youth experience in realising for the first time the racial divide. Other African Americans have described similar experiences. Melba Pattillo Beals, for example, one of the nine African American teenagers that tried to enter Little Rock Central High School, had such experiences at the age of four and five (2007:3-4). Also, see Ellison (1989) and Lawrence III (1987), who starts his article by telling an experience he had as a five-year-old, as the only black child in the classroom.

35 The veil is also ambiguous in other respects because it is used as a symbol of both freedom and repression: see Ahmed (1992). For religious views on seeing and blindness: see Resnik and Curtis 2011:64-65).

How then should we understand the veil of prejudice? Should we model our understanding on Rawls' or Du Bois' use of the veil metaphor? The thread running through the next chapters will precisely be to explore what epistemic model should guide us in understanding the veil of prejudice. So far, what I provisionally present as two epistemic models, are still only metaphorical images. In the following sections, I will introduce two different frameworks: mentalism and pragmatism. The first I associate with the epistemic model underlying Rawls' veil metaphor, the second with the one underlying Du Bois' veil metaphor.³⁶ For how I want to understand the veil, in the end, I reject Rawls' use of it. Instead, I would like to learn from Du Bois' use, especially for its use of the ambiguousness of the veil. But I adapt his veil metaphor in several ways. First, I want to use it in a more abstract sense, not just for issues of racism but for a broad range of issues that involve social prejudices. Next, I want to reverse Du Bois' metaphor in two senses. I want to reverse the perspectives of who is looking through the veil and who is looked at. For the veil of prejudice, I will start from the perspective, not of those who suffer from practices of hidden prejudice, but of those who perform those practices. I also want to reverse Du Bois' diagnosis of 'double consciousness': I will assume that the problem with prejudiced people is that they are *lacking* the perspective of looking at themselves 'through the eyes of others'. For them, therefore, the double-mindedness that discriminated people suffer from has narrowed down to one-mindedness. And what they *see*, what they *understand* of themselves and of what they are doing for other people, may very well present them with a tolerant, non-discriminative picture.

What nowadays seem to be the dominant strands of theorising practices of hidden prejudice are at least Critical Race Theory and social psychological research on 'implicit bias' (which is the social-psychological phrase for 'hidden prejudice'): the first strand focussing on racist practices, often described as structural racism, the second focusing on the psychological roots of those practices. While Critical Race Theory is valuable for its insights into the workings of racism, I will not make use of it to unravel the veil of prejudice. Instead, I will focus on the social psychological research on implicit bias because I am interested in the connection this research has, first, with an underlying mentalist framework (see section 1.4.) and, second, with philosophical liberalism and especially with liberalism's notions of responsible agency (see section 1.5.). My interest in both connections will be critical. Critical Race Theory is also critical of philosophical liberalism³⁷ but not necessarily or not mainly from the perspective I will focus on, namely the framework that underpins notions of responsible agency (again see section 1.4.). It is also not clear whether Critical Race Theory, to the extent that it can be viewed as a

36 Du Bois' use of the metaphor of 'double consciousness' might seem to commit him to a mentalist framework (see, e.g. this chapter, sections 3. and 4.). But his use, I think, is essentially a metaphorical one.

37 See, e.g. Mills (1997).

single approach, draws on mentalist assumptions or instead is basically critical of it.³⁸

Social psychologists suggest the possibility that implicit biases manage to influence people's actions *unconsciously*³⁹, as shown in an experiment performed in Sweden when, during the selection for job interviews, employment recruiters favoured applicants with Swedish names over those with Arab names.⁴⁰ The Implicit Association Test, which I will discuss later in more detail (chapter 3, section 2), is one of the experiments that social psychologists have designed to measure hidden prejudices or, as they call it, 'implicit biases'.

Gaining more insight into hidden prejudices will pave the way, I hope, for understanding how we may succeed, in a morally justified way, in *confronting* practices of hidden prejudice or, as we may now put it metaphorically, *unravelling* the veil of

38 Some authors from the Critical Race Theory seem to work with mentalist assumptions. See, for example, Lawrence III (1987). While he has shown a critical interest in the research on implicit bias – see Lawrence III (2008) and Lawrence III (2015) – his own theory is informed by Freudian notions of the unconscious, which I choose not to focus on (see chapter 3.2.). Also, see Ngo (2016), who shows interest in the research on implicit bias, but favours, drawing on Merlau-Ponty, a phenomenological analysis of racism as a habit. Also, see King (1991), who works with the notion of 'dysconscious racism'. While for this notion she employs mentalist vocabulary, dysconscious racism is conceptually distinct from 'implicit bias' because it is 'not the *absence* of consciousness (that is, not unconsciousness), but an *impaired* consciousness or distorted way of thinking about race as compared to, for example, critical consciousness' (1991:135). My suspicion is that, in the end, she does not even rely on a mentalist framework, because elsewhere she explains that she is 'interested in the sociology of knowledge that functions as ideology' (1992:319), drawing on a tradition of critical analysis to which for example Marx and Habermas belong and which analyses 'consciousness' as ideology. I even think that, as she defines 'dysconscious racism' as 'an uncritical habit of mind' (1991:135), it may even come very close to the pragmatist notion of hidden prejudice that I develop in chapter 6. However, exploring these conceptual connections has not been part of my analysis. Finally, Woods (2017) might criticise my choice of focussing on implicit bias theory instead of Critical Race Theory. In a critical discussion of the research on implicit bias and the enormous attention (popular and academic) it receives, he argues that the insights of this research had already been anticipated by various authors from the Critical Race Theory ever since the 1970s. I do not intend to deny this. But I do claim that the research on implicit bias is more clearly and more systematically underpinned by a *mentalist* framework (see chapter 3, section 2, and chapter 4, section 2), which is the focus of my argument.

39 As I mentioned earlier in this section, from a legal perspective, two forms of discrimination can be distinguished: direct and indirect, as one between intentional discrimination and between discrimination as the effect of otherwise neutral rules. The notion of 'implicit bias', as social psychologists have introduced it, seems to suggest a new category: unconscious, i.e. unintentional, discrimination. Combining the conscious/ unconscious (intentional/ unintentional) distinction with the direct/ indirect distinction seems to open up new options for a more fine-grained analysis of discrimination. A few authors have already touched on or explored combining these distinctions. For example, Rasmussen (2020) uses both distinctions to argue for four instead of two forms of discrimination, connecting implicit bias to direct discrimination as one of those four forms. Interestingly, Holroyd (2017), who limits her discussion of implicit bias to direct discrimination, worries that a connection of implicit bias with direct discrimination will dissolve the distinction between direct and indirect discrimination. A different strategy is suggested by Richard Ford who, in an interview that Mercat-Bruns (2016:66) had with him, comments that the 'implicit bias theory is basically an argument in favor of disparate impact, an argument to make it stronger – and also an argument for changing the way we distribute the burden of proof in intentional discrimination cases, making the law more friendly to plaintiffs'.

40 Rooth (2007).

prejudice, enabling people to understand their practices of prejudice and change them. The set of actions that might confront practices of prejudice includes a wide range of different *types* of actions. It may include, for example, institutional solutions, like creating an institute where people may file their discrimination complaint: such solutions aim at confronting practices of prejudice by providing an opportunity for people to voice their complaints about discrimination. For my research question, I will focus on actions that aim directly at influencing the motivations of prejudiced people and, within this category, on actions that are *non-deliberative*.⁴¹ The reason for focussing on non-deliberative actions, rather than including also deliberative actions, is that a focus on non-deliberative actions will allow me to explore the mentalist assumptions underpinning hidden prejudice and responsible agency (see next sections). I will now briefly consider both types of action and discuss them according to their effectiveness and their moral permissibility.

The standard way of holding responsible would be the deliberative one: we might make claims against people about hidden prejudices they have, argue that these have discriminating effects and appeal to their ideals of equality, hoping that this will persuade them to adjust their practices. Deliberative strategies, by either participating in public debate or adopting a litigation strategy, require high levels of education that may not be available to people who suffer from practices of discrimination, precisely because of those very same practices. The European Union Agency for Fundamental Rights reported, for example, that, in various member states of the European Union, the illiteracy rates among Roma people are persistently higher, sometimes much higher, than for non-Roma people.⁴² But an even more serious problem seems to prevent the effectiveness of deliberative strategies: to *persuade* implicitly biased people into subscribing to ideals of equality would seem to be meaningless as they already endorse them. On the other hand, attempts at convincing them that they are *really* racists (and therefore should reflect more carefully, more attentively, on their implicitly prejudiced actions?) may not be promising either.⁴³ And so, cases of implicit bias, at least some of them, perhaps many, or perhaps even of all of them, suggest that deliberative approaches fail as a way of holding people responsible for practices of hidden prejudice.⁴⁴

What other means do we have for confronting practices of hidden prejudice?

41 The effects of actions are sometimes more commonly discussed in the vocabulary of 'power'. In some sense, 'having power over other people' will be a thread running through my analyses of 'influencing actions', but not one to which a discussion of notions of power in social-political contexts is needed, I think. For the 'messy' conceptual relations between influence and power, see: Zimmerling (2005).

42 European Union Agency for Fundamental Rights (2014).

43 See Blum (2002); Billig (2012).

44 Exploring whether this suggestion is correct is an interesting issue, but one that I choose not to analyse explicitly. I do claim, however, that *if* one is committed to the idea of implicit bias, this will generally imply that people cannot be held responsible for their implicit biases (also see chapter 4, section 2, and chapter 7, section 8) or if this is denied, that it would make no sense to adopt a deliberative approach in holding them responsible.

Whenever deliberation as a ‘game of giving and asking for reasons’⁴⁵ appears not to work in influencing people, then perhaps another option might be *non-deliberative* actions, some type of actions that influence people without expressing reasons. This, of course, is merely a negative definition. What ‘non-deliberative’ does mean, more precisely and in a positive sense, needs yet to be explored and will be part of my attempt to answer the research question. At least, I will assume that the distinctive feature of such actions is that they influence people in a non-deliberative way.

One possible example of a non-deliberative action has been indicated by scholars who in the last decades have begun to explore various ways in which hidden prejudice (‘implicit bias’) may be reduced. One study, for example, took as a starting point the finding, as previous studies had established, that ‘compared to explicit prejudice, implicit prejudice involves a stronger emotional component’. Translated in biological terms, it means that when people whose implicit biases are activated, for example, when they view the picture of people who are the object of their implicit biases, their brains show an increased amygdala activity. The researchers then succeeded in showing that administering propranolol, a beta-blocker, ‘significantly reduced implicit but not explicit racial bias’ (Terbeck et al. 2012:422) because ‘propranolol can reduce amygdala response to visual stimuli associated with negative emotions such as fear and anger’ (id.).⁴⁶ In a few public media, it was even hailed as an ‘anti-prejudice pill’.⁴⁷ What is interesting about the administration of an anti-prejudice pill is that it influences people in a *non-deliberative* way.⁴⁸ The search for an anti-prejudice pill is, in fact, consistent with various interesting experiments that psychologists, neuroscientists and other scholars have conducted in recent decades to find out how the brain works. In these experiments, they have been able to induce people to perform certain actions without taking recourse to either coercion or deliberation, without even any deliberate attempts to change their subjects’ minds. They did this by *non-deliberatively influencing* them, i.e. by activating (‘priming’) certain dispositions in a way that bypasses their subject’s deliberative reflection.⁴⁹ To put it differently: *what is non-deliberative about such actions is that the influence they exert on people’s minds – their feelings, their motivations, their beliefs – is invisible to their introspective abilities.*

One way of qualifying such non-deliberative actions would be to describe them as manipulation. One of the paradigm cases of manipulation is familiar to the general public: hypnosis, one of the early techniques, reaching back at least to the nineteenth

45 Brandom frequently uses this phrase which he ascribes to Sellars.

46 Also, see Gallate et al. (2011).

47 See <https://nypost.com/2012/03/12/a-chill-pill-for-racists/> and <https://www.thedoctorstv.com/videos/anti-prejudice-pill>.

48 I have chosen the administration of an anti-prejudice pill as exemplary for a mentalist approach. But it also supports other types of interventions. See chapter 7, section 8.

49 See Brasil-Neto et al. (1992); Wegner and Wheatley (1999). The second case is discussed in chapter 4, section 5.

century, that psychologists have used for tapping into the deeper layers of people's unconsciousness.⁵⁰ Today a variety of techniques that seem to work in a similar way and which is better known as 'nudging' is already being explored and applied by policymakers, for example, to steer citizen's choices in directions that will improve their health.⁵¹ If manipulation is understood according to these examples, this introduces what seems to be the equivalent of the veil of prejudice – the veil of manipulation – because here, too, a boundary exists between people's self-knowledge and their self-ignorance. But there is a slight difference: while cases of hidden prejudice follow the pattern of a hidden motivation that produces an action, the pattern in cases of manipulation is one of a hidden action that produces a motivation. If non-deliberative actions are manipulative in this sense, they may therefore be effective in reducing practices of hidden prejudice. Letting people swallow an anti-prejudice pill would provide such an example.

This description of manipulation suggests that manipulating people is a good way of holding them responsible for their practices of hidden prejudice. But manipulation has, of course, a bad reputation – not just in popular culture but also in political theory – because it is supposed to violate people's responsible agency. It may explain why manipulation has not yet been explored as a possibly justified means of holding people responsible for their practices of injustice.⁵² In recent years however various philosophers have argued for definitions of manipulation that either separate its technique from its moral assessment or allow room for special cases in which manipulation is permissible. This might open new perspectives for exploring forms of actions that are neither violent nor deliberative but may still be both effective and responsible in *confronting prejudiced behaviour*. Nevertheless, in focussing on the category of *non-deliberative* actions, the challenge will be to figure out whether, or to what extent, non-deliberative actions are manipulative (in the bad sense) or whether they can be manipulative in perhaps some neutral sense.

Now that I have explained the moral perspective that I take on confronting discrimination and how I limit my analysis to hidden prejudices and non-deliberative

50 On the early uses of hypnosis in psychology: see, e.g. Hammond (2014).

51 By now, there is extensive literature on nudging, inspired largely by Thaler and Sunstein (2008). One of the issues it treats is whether nudging is a form of manipulation. It also deals, different from the standard literature on manipulation (see chapter 2, note 68), with the issue of whether nudging is allowed, as a form of paternalism, to foster to improve the welfare of people that are nudged.

52 Literature on manipulation as an instrument to achieve morally good ends seems scarce. Only sometimes do theorists make passing comments on this possibility. See, e.g. Feinberg who describes the situation that 'manipulative techniques are used to open a person's options and thus increase his freedom on balance, but without his consent' (1986:67) but disapproves of such manipulation in which a person is 'manipulated into freedom' (id.). Also, see Greenspan, who describes cases of 'less self-interested manipulation' (e.g. someone cheering up a melancholic friend), which she calls 'examples of a kind of paternalism, but far outside the legal contexts in which we object to paternalism' (2003:155-156). I have found no analyses of manipulation as a possibly justified form of opposing injustice. What comes closest are Bowers' comments on the possibility of manipulative strategies 'as a low-key way of fighting back at a system which has deprived the prisoner of normal freedom' (2003:330).

actions, it is time to formulate my research question. It will be as follows:

What are the conditions for non-deliberative actions to count as holding responsible those people who perform practices of hidden prejudice?

1.3. Methodological approach and research aims

In answering the research question, my methodological approach will consist of making the following moves:

1. I will treat the research question as requiring that *practices of hidden prejudice* and *non-deliberative actions* are analysed in terms of *responsible agency* (i.e. of those who perform practices of hidden prejudice and are affected by non-deliberative action).
2. I will focus on analysing responsible agency in *epistemic* terms or in terms of, as I will call it, *epistemic autonomy*.⁵³
3. I will analyse epistemic autonomy in terms of *rationality*, i.e. as claims for knowledge.
4. I will analyse responsible agency also in terms of *freedom*.
5. I will analyse my *central notions* (see figure 1) from two different *frameworks* – mentalism and pragmatism – that I will discuss and contrast at *different levels of abstraction*, each of which – either the concrete level of specific notions or the abstract level of the framework itself – requires its own methodological approach.

In this section, I will first briefly comment on my various methodological moves, including a few remarks on my use of terminology, and then state my research aims.

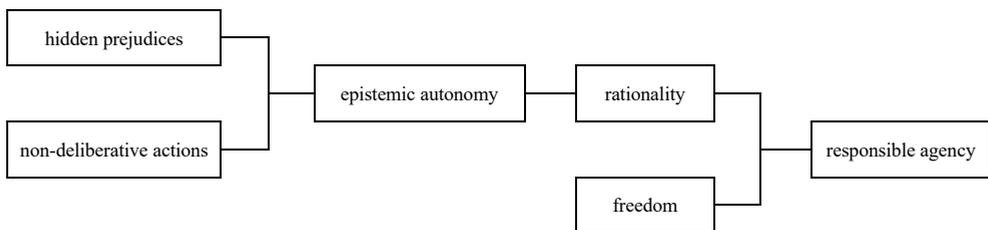


Figure 1: Central notions.

⁵³ The term ‘epistemic autonomy’ seems relatively new in the philosophical vocabulary and is still rarely used. But it is used in, for example, discussions within social epistemology to explore to what extent an individual should, or can, rely on other people’s testimony to the truth of certain beliefs. To the extent that this is denied, autonomy is defined as a form of self-reliance. See, e.g. Elizabeth Fricker (2006). My use of the term ‘epistemic autonomy’ is motivated, however, by different concerns (see, e.g. chapter 4, section 4).

In explaining my methodological approach, I will start with my first four moves. In the previous section, I introduced hidden prejudice and non-deliberative action. Holding people responsible for their practices of hidden prejudice seems to presuppose that they are responsible, somehow, for their hidden prejudice. It suggests that an act of holding them responsible is permissible *if* their action of hidden prejudice is rooted, *in the appropriate sense*, in their responsible agency. The phenomenon of non-deliberative action, on the other hand, suggests that an act of holding people responsible is permissible if that act itself *respects* (or: does not violate) the responsible agency of those who are held responsible. We might provisionally call the first condition *the appropriate link condition* and the second *the respect condition*.

It would seem then that the assessments of both conditions are separate issues. But this is not entirely correct. On closer view, it appears that both hidden prejudice and non-deliberative action invite an understanding of responsible agency in specifically *epistemic* terms. As we have seen in the previous section, they raise questions on self-knowledge and self-ignorance and where, or how, to draw the boundary between them. It suggests that we should focus not so much on responsible agency as such but on the epistemic dimension of responsible agency. The research question, therefore, as I choose to analyse it, does not turn on two different *issues* – i.e. hidden prejudice and non-deliberative action – but on two different *routes* to basically one issue: *epistemic autonomy*. It might provide one important condition for responsible agency and therefore also for non-deliberative actions to count as holding people responsible for their practices of hidden prejudice. If hidden prejudices or non-deliberative actions interfered with people's epistemic autonomy, they would not be responsible for their hidden prejudices, and non-deliberative actions would be impermissible as acts of holding responsible. My earlier example of the anti-prejudice pill illustrates how *one* action – letting people swallow an anti-prejudice pill, especially when it is done secretly – may interfere with people's epistemic autonomy in two distinct ways.

The notion of epistemic autonomy, of course, raises the question of what it means. Only then can we explore what it means that people's epistemic autonomy is interfered with. One possible strategy would be to claim that people have epistemic autonomy if they have self-knowledge of the beliefs (even false ones such as prejudices) that somehow structure certain motivations for actions, or conversely, that they do not have epistemic autonomy if they are self-ignorant about such beliefs. But this strategy would miss the point about autonomy because it might be that people just 'happen' to be self-ignorant, or not, about their prejudices. Instead, we should question the boundary itself – what I have already called, in a metaphorical sense, the veil – between self-knowledge and self-ignorance. What is people's basic attitude towards their beliefs? Are they basically 'passive' towards their belief? Do they only 'happen' to have the beliefs they have, on whichever side of the veil? Or is it possible to describe the beliefs they have, on both sides of the veil, as the result of some *autonomous* epistemic process of belief formation? If that made

sense, it would introduce the possibility that people are even responsible for the beliefs they have become ignorant about. It is not enough, therefore, to observe some veil hanging between people's actual self-knowledge and actual self-ignorance at some point in their lives. Epistemic autonomy as a norm also requires that we explore the origin of this actual veil. I choose this strategy, and I choose to analyse epistemic autonomy in terms of rationality, exploring whether epistemic autonomy can be identified with some form of rationality.

Responsible agency requires not only epistemic autonomy but also *freedom*, not in any political sense but in a sense opposed to determinism. I choose to introduce this assumption as my next methodological move because it will allow me to explore the conceptual relations between rationality, as the criterion for epistemic autonomy, and freedom. And this, in its turn, will allow me to highlight the essential differences between the two different frameworks from which I will try to answer my research question. This brings me to explaining the final move of my methodological approach.

Contrasting different views on an issue may consist of analyses at different levels of abstraction. It may analyse them as consisting only of different notions for the same 'thing'. Such rival views, for all their differences on these notions, may still share many assumptions and implications if they are set within the same theory. Views may also be contrasted as being different theories. Even then, they will still share many assumptions or implications, although not as many as the views that differ on notions. For the argument in the next chapters, I will assume that mentalism and pragmatism are views that differ from each other in defining a framework. As I understand a 'framework', it is not so much a specific theory but a conceptual architecture that functions as a deep background to a family of theories. If frameworks are being contrasted, they will therefore share fewer assumptions than whenever either notions or theories are contrasted. What is different between these various levels of contrasting is, therefore, the level of abstraction: the extent to which either the notions, the theories or the frameworks that are being compared share certain assumptions or implications.⁵⁴

My analyses, at whatever level of abstraction, will be mostly conceptual. In discussing some notions, I have nevertheless provided some historical context or at least tried to reconstruct, to some extent, the origin of certain notions or debates, not because this is an aim in itself but because it supports certain claims I make in the context of a basically conceptual analysis.

My methodological approach will not be the same for each level of abstraction as I have different aims at each level. At the level of notions, I will be particularly concerned with conceptual clarification, consisting of a detailed, close reading of relevant literature,

54 Distinguishing, as I do here, between notions, theories and frameworks as different levels of abstraction is only meant as an explanatory, heuristic device. In a sense, it is a variation of Rawls's idea to distinguish between concepts and conceptions. I do not mean to suggest that we can only distinguish three levels of comparison. The idea of a 'family of theories' is, of course, indebted to Wittgenstein. The notion of framework is related to Kuhn's notion of a paradigm.

as my aim would be to construct or argue for a specific notion. At the level of frameworks, my aim will be to make plausible, especially regarding mentalism, that it is a framework at all, distinctive from pragmatism and one that provides the background for the standard debates on, for example, responsible agency, freedom and other notions that are central to my argument. For this reason, I will be concerned, at this level, with a more diagnostic reading of relevant literature and, to some extent, with reconstructing the history of some debates or some notions. As a result, I deal with relevant literature in different ways, sometimes from a microscopic perspective (conceptual clarification), other times more distanced, from a telescopic perspective (diagnostic reading).

The challenge of contrasting two different frameworks is to stay clear on what terminology belongs to one or the other framework. The two frameworks sometimes use the same terminology for what are different notions, as in the case of the adjective ‘implicit’.⁵⁵ This may complicate an adequate understanding of important differences in their conceptual backgrounds. Although I hope this will become sufficiently clear to my readers as I go along with the argument, here I want to make explicit that, as umbrella terms that are supposed to be relatively neutral between both frameworks, I have chosen at least the following: responsible agency, hidden prejudice, non-deliberative influencing, epistemic conditions, self-knowledge and epistemic dispositions (this last phrase is used for a definition of manipulation that is supposed to be neutral to both frameworks: see chapter 2, section 3). As we will see, those working within a mentalist framework often use specific terms that I will treat as the mentalist versions of some of my umbrella terms. They will sometimes use, for example, ‘implicit bias’ (for hidden prejudice), ‘manipulation’ (for non-deliberative, influencing action), ‘cognitive conditions’ (for epistemic conditions) or ‘introspection’ (for the process of self-knowledge). As the pragmatist framework is not as fully developed or yet recognised as a competing framework, it lacks in many ways its own terminology. Whenever I discuss the pragmatist framework, I will take recourse to the umbrella terms, but I will use them in a pragmatist sense.

My research aims are as follows. First, I want to show that *if* my central notions were understood from within a *mentalist* framework, this would leave *no conceptual room* for understanding non-deliberative actions to count as holding responsible. The reason would be that the *epistemic conditions* under which non-deliberative actions would count as holding responsible simply do not exist. Non-deliberative actions cannot ever, by definition, count as holding responsible, first because people are not responsible for their hidden prejudices, second because influencing non-deliberative actions do not respect people’s responsible agency. In the second place, my research aim is to show that we *can* make conceptual room for understanding non-deliberative actions as holding responsible if only my central notions are understood according to a *pragmatist* framework, (1) *if* hidden prejudices are understood as involving a boundary between

55 This adjective is introduced and discussed in a mentalist sense in chapters 3 and 4, in a pragmatist sense in chapter 5.

ignorance and knowledge about one's prejudices (1a) that is ambiguous and as admitting of degrees, rather than as clear-cut, and (1b) that is primarily a social boundary, rather than an individual (intra-psychic) one; (2) *if* non-deliberative, influencing actions are understood as appealing to people's *rational capacities*, without denying that they operate in a non-deliberative way; and finally (3) *if* responsible agency is understood as a social status rather than as an individual state. Achieving my research aims would mean that a pragmatist framework can be used, where the mentalist framework cannot, to answer the research question, i.e. to explicate the conditions for non-deliberative actions to count as holding responsible people with hidden prejudices.

1.4. Two frameworks: mentalism and pragmatism

In using these two frameworks, mentalism and pragmatism (in the version that I will argue for), as the background for my discussions, my research aim is *not* to provide an extensive, detailed, comparative analysis of either their historical roots or their conceptual differences. This is beyond the scope of my argument. In the following chapters, I will analyse their differences only to the extent relevant to my argument. It may nevertheless be helpful to my readers to be introduced, beforehand, to both frameworks and have some idea of their differences. In this section, I will therefore provide a brief outline of their basic conceptual structure. I will also explain my choice for these two frameworks. Finally, I will suggest, again very briefly, some historical background to my assumption that both frameworks represent two very different proposals to understanding responsible agency.

Mentalism, to begin with, offers a set of assumptions on the human mind.⁵⁶ Its basic strategy consists of analysing the mind in terms of mental states and qualifying these states as either conscious or unconscious.⁵⁷ Knowledge is understood in terms of a subset of these

56 What I mean with 'mentalism' corresponds to what Descombes targets and criticises in his (1995). Also see Mittelstraß (1995); Thümmel (2019). Arthur W. Collins (1978). As a synonym, also 'cognitivism' might be used: while Descombes (1995) uses the French word 'mentalisme', his translator, Stephen Adam Schwartz, chose for the English translation (2001) to add the subtitle 'A critique of cognitivism'.

57 The notions of 'conscious' and 'unconscious' that I have in mind here are to be understood as ontological categories, i.e. as the mental counterparts of physical processes. It follows the dualistic model that Descartes is supposed to have introduced (see chapter 2, section 6, especially note 108). This does not mean that ever since Descartes, the conscious/unconscious vocabulary has been interpreted as an ontological one. In nineteenth-century German philosophy, for example, it has been viewed as a metaphysical category (Gödde 1999; Völmicke 2005). The history of this vocabulary is, therefore, a complex one: see early studies, like those of Grau (1916), Whyte (1960), Ellenberger (1970) and Klein (1977), which have been supplemented and corrected by more recent studies, like those of Davies (2008), Nicholls and Liebscher (2010b) and Ffytche (2012). Of course, the best-known notion of the unconscious is Freud's. Freud's notion of the unconscious. However, for my argument, I will focus on the notion of the (un)conscious that social psychologists use (see chapter 4, section 2). Apart from various reasons for making this choice (see chapter 3, section 2; also, chapter 4, section 2), Freud's notion is also more elusive than the social psychologists' notion of the unconscious. Although it has been interpreted as an ontological category, this interpretation has been disputed. See, e.g. Woody and Phillips (1995).

mental states: as beliefs. Within the mentalist framework, the notion of beliefs is based on *representationalist* semantics. It means that beliefs are related to certain states of affairs in the world as representations of those states of affairs.⁵⁸ However, not all beliefs are necessarily correct beliefs. According to one tradition within the mentalist tradition, whether you have a correct belief is whether it is a justified true belief.⁵⁹ Therefore, the correctness of a belief depends in part on whether you have reason to believe a certain state of affairs is true. While self-knowledge would seem different from knowledge about the world, it is basically modelled on beliefs people have about the world, except that having beliefs about yourself are representations of what goes on in your mind. Self-knowledge within the mentalist framework is understood as introspection, a process of becoming conscious of one's own mental states, which is sometimes claimed to be a kind of direct knowledge as opposed to the indirect knowledge we have about the world. And of course, here, as in the case of our beliefs about the world, what decides the correctness of our beliefs, is whether we are justified in having these beliefs about ourselves. Rationality plays the role of deciding between correct and incorrect beliefs (about either the world or ourselves) as a feature of the processing of (epistemic) mental states.

The reason for choosing mentalism is that it has been, and still is, the dominant view on the human mind in at least Anglo-American philosophy: it underlies a broad range of debates within empiricist and analytic philosophy⁶⁰ and, sometimes too, moral and political philosophy. Via philosophy of mind, it also underlies the theoretical framework of natural-scientific psychology and, more specifically, of social psychological research on 'implicit bias' (and also 'nudging') because the 'implicit' nature of implicit biases is understood as what remains hidden from people's introspection, i.e. as 'unconscious'. And finally, it also has deeply shaped various conceptions of responsible agency, even to the extent that philosophers have accepted the mentalist vocabulary as

58 Many mentalist claims about how the mind works have been criticised, but usually from *within* the mentalist framework. The mentalist framework as such is not so often identified, discussed or criticised. Mentalism is a modern framework whose origins can be traced back to early modern (epistemological) theory. For critical, historical reconstructions of the origins and development of the mentalist, representational epistemology: see Foucault (1966) and Rorty (1979). On the difference between both reconstructions, Rabinow (1986/ 1996:34) remarks that '[i]nstead of treating the problem of representations as specific to the history of ideas [as Rorty does], Foucault treats it as a more general cultural concern'. Toulmin, too, has written several critical, partly historical articles on the mentalist tradition, but with a focus on psychoanalysis. See, e.g. Toulmin (1986). Foucault also resisted the internalist gaze of the mentalist tradition, but he too focuses on psychoanalysis. For a discussion: see Whitebook (2005). For a historical account of mentalism from the perspective of psychology as a new discipline: see Richards (1992). In recent years various Anglo-American philosophers have been exploring the (continental) phenomenological tradition to counterbalance the dominance of the mentalist framework: see, e.g. Gallagher and Zahavi (2012). For a recent, interesting criticism from a French philosopher: see Descombes (1995). The rise of the mentalist framework should be understood, I think, within a broader development in which the human self was increasingly understood in subjectivist terms: see Taylor (1989). More on the mentalist tradition: see section 2.6.

59 See Brandom (1998).

60 See Manson (2000).

the most natural way to analyse responsible agency. Whatever the differences between these conceptions, their common core meaning can be called the standard notion of responsible agency: people can be held responsible for their actions if they are conscious of the motivations for their actions.

With its vocabulary of consciousness, mentalism was not originally intended as a framework supporting a naturalist programme.⁶¹ However, it turned out to be a useful tool for a strategy to naturalise people's agency in due time. That, at least, is the hypothesis underpinning my argument. And therefore, whenever I talk about the *mentalist* framework, I usually mean to talk about the *mentalist-naturalist* framework.

Very much simplified, one could say that the naturalist strategy, employing the mentalist framework, consists of the following four steps. The first step is to understand *mental states* in *empirical* terms. In this way, they are basically understood as states that somehow have their place in the human head (the brain). It is open to debate how exactly we should understand mental states as having their place in the head, or perhaps more precisely, in the brain. Nowadays, the debate has moved towards a neurologically informed interpretation. Whatever the exact claim may be, the essential move is to understand mental states as ontological states, as states therefore that in principle might be said to have a certain place and time in the world. For that reason, they are necessarily individual states, states that belong to a specific individual's brain. A second step, which focuses on understanding knowledge in terms of mental states, is to understand *beliefs* too as empirical, i.e. mental states that have their place within a specific individual's brain.⁶² This step is often followed by a third step: choosing a strategy that seeks to demonstrate that beliefs should be understood as *internal*, i.e. *intra-psychic* phenomena. This strategy consists, therefore, in tying meaning to the empirical circumstances of the brain.

So far, these three steps do not yet, by themselves, imply anything about responsible agency. This naturalist strategy might concede that people may have epistemic autonomy of some sort but claim that this does not imply responsible agency. The strategy used to naturalise people's knowledge about the world would be to assume that individuals have a passive role towards their beliefs: the acquisition of knowledge is determined by how the senses work biologically (neurologically).⁶³ There is nothing we can change about them. However, this move does not seem to be available to naturalise people's agency as this seems to require, by definition, an active role of people: how can we understand

61 On mentalism as supporting a naturalist programme: see section 4.2.

62 The problem that has puzzled philosophers of mind has been, of course, how to understand the relationship between mental states (both conscious and unconscious states) and the brain.

63 Underlying the idea that people, as knowers, have a passive role is the metaphor that people, as knowers, are mirrors of the world. They mere *reflect* what they see and observe. The metaphor can also be used for knowledge of subjectivity: in that case, the observing eye of the knower reflects on, introspects, on her own 'inner world'. Various others have commented on this kind of 'ocular metaphors', as Rorty (1979/ 2009:13) described them, that are typical for the mentalist framework. Interestingly, these metaphors tend to single out 'seeing' as the most important of the senses and as providing the model for acquiring knowledge.

people as being passive towards their own actions? Especially at this point, the mentalist framework has proven to be very fruitful in elaborating the naturalist approach. Theorists working within the mentalist framework will usually acknowledge two basic models for explaining the motivational pattern ‘motivation A leads to action B’. It may be explained as belonging to either the ‘space of reasons’ – i.e. the *reasons* model, which implies: ‘A is a reason for B’ – or the ‘space of causes’ – i.e. the *causes* model, which implies: ‘A causes B’. The space of reasons and the space of causes are usually treated to correspond with the space of *conscious* and *unconscious* mental states. When the acting person is conscious of the reason (‘A’) which motivates her to perform some action, she is supposed to have, in this respect, responsible agency. Therefore, the fourth step in a naturalist strategy would be, very generally, to explain the reasons model in terms of the causes model as more basic.

Mentalism presents, in my view, a deeply problematic approach to responsible agency. For this reason, I have chosen to explore yet another framework and contrast it with the mentalist framework. I will take recourse to the philosophical tradition of pragmatism and, more specifically, rely on Robert Brandom’s appropriation of that tradition.⁶⁴ Whenever I talk about the *pragmatist* framework, I will usually mean a *Brandomian-pragmatist* framework.

Pragmatism as such is not an obvious choice. Although pragmatism has more often been used as a source for criticising mentalism, it does not necessarily provide an alternative to naturalism.⁶⁵ Furthermore, pragmatism, as a philosophical framework, offers a certain approach to understanding knowledge, language, and meaning rather than a political view or theory.⁶⁶ Even now, when various prominent philosophers are usually recognised as working within a pragmatist tradition,⁶⁷ pragmatism is still not associated, within social and political philosophy, with a distinctive approach to analysing social and political problems.⁶⁸ Various philosophers are nevertheless exploring the potential

64 The beginnings of this tradition are usually traced to the American pragmatists (Peirce, James, Dewey and Mead). For a general account of pragmatism: see Legg and Hookway (2019) and, more specifically, addressing its contribution to political theory, see Goldman (2015). Pragmatism, as a philosophical position, is also represented within continental philosophy: see Vossenkuhl (1974); Schneider (1975); Gethmann (1979). For an interesting discussion of Brandom’s attempt to explain meaning as use as compared with similar attempts of some of these German-speaking philosophers from the Erlangen/Constance school: see Jäger (2001).

65 Radical behaviourism, for example, having its roots in pragmatism, opposes mentalism but has a thoroughly naturalist outlook in embracing determinism and denying free will. See Lattal and Laipple (2003); Moore (2005) and Moxley (2004).

66 See Goldman (2015).

67 Richard Rorty has occasionally been referred to as a liberal pragmatist. Habermas and Putnam are usually recognised as adopting a pragmatist approach to political philosophy. See Legg and Hookway (2019).

68 Discussions in handbooks of political philosophy on pragmatism are virtually missing, as in, for example, Kymlicka (2002) and Estlund (2012). Systematic studies on political theory and pragmatism are scarce. For exceptions: see Festenstein (1997). It has recently been questioned whether pragmatism, as a tradition of theories on knowledge, language, and meaning, is suited to normative political theory. See Erman and Möller (2014) and (2014). For a critical response: see Fossen (2019).

of a pragmatist outlook for political-normative theory. They are especially focussing on deliberation (Misak 2000; Talisse 2005), the rule of law (Posner 2003) and legitimacy (Fossen 2011). Once again, their choices in using pragmatism for political-normative theory are not of help to my argument. My interest in the pragmatist tradition for theorising political problems differs from theirs as I am primarily concerned with *non-deliberative* actions, i.e. actions whose non-deliberative effects on people are described, from a mentalist perspective at least, as what they are *unconscious* of. While pragmatist-political theorists usually draw for their inspiration on the writings of, for example, Dewey, the late Wittgenstein or Rorty, I choose the writings of Robert Brandom. They have been devoted mostly to issues of language and logic, at least not to those of normative political theory. They offer, however, a particularly interesting, although complex, theory of practice as normative and social. What scarcely has been explored is its potential for supporting normative political notions.⁶⁹

I will also draw on the views of a sociologist who might be qualified as a pragmatist: Bourdieu. Brandom and Bourdieu seem to be an unlikely pair of allies. They start from different intellectual traditions (Bourdieu from phenomenology and structuralism; Brandom from American pragmatism and analytical philosophy), from different disciplines (respectively sociology and philosophy) and from different engagements: Bourdieu is primarily interested in understanding non-linguistic practices, Brandom in the sociality of linguistic practices. But at heart, their projects have the same ambition: understanding human beings in the first place from the perspective of intersubjectivity, i.e. as social beings. They also get there by making the same choices: refusing to understand human beings as mere mechanical beings (however complex they are) and refusing to understand them as rational-intellectual beings. Because of their differences, their approaches have developed along different lines, in different vocabularies and in polemics with different adversaries.⁷⁰ My aim has not been to integrate the views of Bourdieu and Brandom. I have chosen Robert Brandom's writings as my main starting point, and I have chosen Bourdieu's views to challenge mentalist conceptions (chapter 2) or to tease out some notions that Brandom's normative pragmatics remain silent about (chapters 6). In these cases, their differences prove fruitful for my discussion.

One of the central themes of Brandom's normative pragmatics is understanding people as knowers (rational beings). A dominant tradition in philosophy identifies rational agency with the ability for deliberation (deliberative rationality). On the external, practical side, this is expressed in the view that citizens, to participate in the public debate, should be able to articulate their views and deliberate with other citizens about each other's views.⁷¹ On the internal, theoretical side, it is expressed in the view

69 An exception is Fossen (2011) and (2014).

70 Either the affinities or differences between Brandom and Bourdieu have scarcely been explored. The only (concise) examination I could find is by Kibble (2014). Others, writing on either Brandom or Bourdieu, have made only occasional observations on their affinities.

71 On the relation between autonomy and deliberation: see Crittenden (2002).

‘that our way of political life is free and rational only if it is founded on some form or other of critical reflection’ (Tully 2003:17). Liberal philosophers seem to be committed too, generally, to the ideal of *deliberative rationality*.⁷² Citizens are supposed to have this ability to develop their personal views on societal issues, but also what it means, for them, to live a free, autonomous life.⁷³

In contrast to this view on rationality as essentially deliberative, Brandom claims a more basic rationality for human practices, namely rationality as *comprehensibility*, more specifically: *inferential* rationality. People are *claim-makers*, not just in the limited sense (to which I think, for example, Rawls is committed) of making claims about what constitutes a good life, about who they are, about their interests, about what they value, but a much broader range of claims. Rationality is inferential because people understand how their claims make sense as the premises or conclusions of certain inferences. Although making claims and making inferences is usually thought of in a linguistic sense, as reasoning, as in a debate with other people, the more basic rationality is constituted by the practices people perform without making their claims or inferences explicit. These practices can be understood as displaying rationality, according to the Brandomian-pragmatist approach, even without the need to assume that people deliberate in some interior, mental sense as an inner conversation. Standard human actions should be understood as non-deliberative yet rational.

Brandom argues that inferential rationality is a social practice in which all participants authorise other people’s claims and take responsibility for their own claims. Whoever authorises herself for making a claim is responsible for giving reasons when asked for it. The other way round, whoever does not hold another responsible for claims she makes, authorises her to make that claim. The act of authorisation is not something that the individual is able to perform on her own, outside a social context. Authority is always linked to responsibility.

As the previous outlines indicate, the differences between both frameworks cut deep. They have very different notions of responsible agency because they start from very different views on how to analyse rationality and freedom as conditions of responsible agency. While mentalism assumes that rationality and freedom are distinct issues, pragmatism treats them as intimately connected and perhaps even identical: freedom is a matter of rationality. For mentalism, rationality is primarily something that is somehow a *thought process*, while for pragmatism, it is primarily something that is *done*, or, as Swindal puts it in a slightly

72 Tully (2003) claims this is the widespread view among contemporary political philosophers. But he is critical of this view. He presents Jürgen Habermas and Charles Taylor as representatives for this view and discusses them critically, drawing his arguments from Wittgenstein’s *Philosophical Investigations*.

73 The liberal ideal of deliberative rationality has been criticised from various perspectives. It conceals, for example, assumptions about how the public debate functions in ways that are too demanding in non-ideal conditions: ‘[a]ctual situations of discussion often do not open themselves equally to all ways of making claims and giving reasons. Many people feel intimidated by the implicit requirements of public speaking’ (Iris Marion Young 2000:39). Also, see Walzer (1989); Sanders (1997).

different way, that for mentalism, the ‘vehicle of thought’ is *(un)consciousness* while for the pragmatist framework, as Swindal puts it, ‘it is action, rather than consciousness, that is the vehicle of thought’ (2012:13). The mentalist framework distinguishes between conscious and unconscious beliefs. Instead, the pragmatist framework starts from the distinction between knowing-that and knowing-how (in Brandom’s vocabulary: respectively explicit and implicit), but the distinction operates in a radically different way.

But the differences cut even deeper. Mentalism is a framework for understanding people’s practices, including their discriminatory practices, in terms of *internalist*, i.e. *intra*-subjective features of people’s mind/body. As a result, rationality is understood as the primary feature of an individual (brain). Mentalism is therefore committed to a methodological, perhaps even ontological, individualism. Pragmatism, however, understands people’s practices in terms of *externalist*, i.e. *inter*-subjective features: rationality is the primary feature of the practices of a community of people. What is rational is something to be assessed in relations between individuals. The implications of this pragmatist idea are far-reaching. It would mean, for example, that assessing a person’s responsible agency is not exclusively a matter of whatever a person, in fact, knows about herself, about her situation, about the world, but also depends on what others know about her, about her situation and about the world. In other words: understanding what responsible agency means depends to some extent on what *holding responsible* means.

The reason for these differences is that mentalism and pragmatism (at least in the Brandomian version that I will argue for) represent two different approaches to understanding agency, respectively what Lance and White (2007) call a ‘metaphysical’ and a ‘stance’ approach.⁷⁴ While the first approach ‘begins by asking what a person, or agent, or subject *is*’ (2007:2), a stance approach asks a very different question, namely

74 Historically, these two approaches can be recognised as emerging in early modern discussions that were concerned with designing a new notion of human actions against the background of ‘[p]olitical and social changes and the advancement of science’ (McKeon 1957/ 1990:72). As McKeon explains, the ‘major alternatives’ for developing new views on responsible agency ‘turned on a choice between accountability and imputation’ (1957/ 1990:72). Their disagreement is on how to understand the moral nature of human actions in relation to freedom and necessity. The second alternative (imputation), which fits with the ‘stance’ approach, chose for a two-stage explanatory strategy: it assumed that we first need an analysis of whether human action is free (as opposed to being determined by scientific laws of nature) and that only then we can move forward to morally assess the human source. The imputation tradition fits with the ‘stance’ approach. For both, the basic strategy is an ‘ascriptive’ one: for assessing whether actions are either ‘free’ or ‘morally good’ is to articulate what it means to impute (ascribe, attribute) ‘free’ or ‘morally good’ to an event. The first alternative (accountability), however, chose for a one-state explanatory strategy by assuming that ‘human actions are determined by a causality or necessity similar to that which determines physical change’ (1957/ 1990:72), but in a sense that this necessity is a *moral* necessity, as distinct from a *physical* necessity, i.e. a necessity that already includes the moral qualification of actions. The ‘disenchantment’ of the sciences, however, made the accountability tradition vulnerable to mechanistic views on responsible agency. Although I will not further specify this specific claim, this ‘historical’ diagnosis underlies my argument against the mentalist framework in chapter 5.

‘what it is *to take something to be* a person/free agent/subject’ (2007:2).⁷⁵ In explaining the stance approach, Lance and White have drawn too on Brandom. As my argument is specifically informed by the problem of practices of hidden prejudice, in the next chapters, I choose my own strategies of explaining and contrasting both approaches.

1.5. Context and relevance

The broad context within which I will explore my research question is liberalism as a political philosophy. More narrowly, the context is one in which liberalism is informed by scientific insights into responsible agency, specifically those of social psychology and social neuroscience.⁷⁶ Liberal philosophers, at least those working within a largely Anglo-American tradition,⁷⁷ have inherited their notion of responsible agency from an empiricist, mentalist tradition.⁷⁸ In this section, I will briefly comment on these contexts and the relevance of my argument for philosophical liberalism.

Responsible agency is or at least should be an important notion to liberalism as a political philosophy.⁷⁹ Its importance may perhaps be illustrated by a brief discussion of

75 Lance and White (2007) compare the stance approach to what Strawson (1962) has called the ‘participant’ stance. Other philosophers they mention as representative of the stance approach are Dennett (no reference mentioned, but it might be 1987), McDowell (1996) and Darwall (2006). Historically, the stance approach covers the work of authors as different as, for example, Pufendorf and Hart. It is beyond the scope of my argument to provide a precise historical-analytical or comparative analysis. But I will suggest a few sources. Historically the idea of an ascriptive, *social* model can be traced back to Pufendorf (1660), who can be said to have started the early modern (largely German) tradition of ‘imputation’ (which I consider as more or less a synonym for ‘ascription’, the term I choose here). Also see Hruschka (1986); Kobusch (1993); Hunter (2004). In Anglo-American philosophy, it is probably H.L.A. Hart whose views fit within the ascriptive model (see Hart 1948 – 1949). To my knowledge, notions of ‘free will’, ‘autonomy’ or ‘moral responsibility’, within Anglo-American debates are not usually theorised following an ascriptive approach. An exception is perhaps Sneddon (2006), who draws on Hart’s views. Also, see Anderson (2014).

76 In an interview that Mercat-Bruns (2016:65) had with him, Richard Ford suggests that ‘We, in the modern, technologically advanced societies of the West – and certainly in the United States – tend to put a lot of stock in technical, empirical studies, even when the question that we are ultimately trying to resolve isn’t an empirical question’, and then adds, commenting on the research on implicit bias: ‘That concerns me. The implicit bias literature feeds into a fetish of social science’.

77 Here I use ‘Anglo-American tradition’ as a diagnostic term to help identify, but only very roughly, a ‘family’ of philosophical theories and debates on liberalism that are mainly indebted to or influenced by liberal philosophers that have been working in Anglo-American countries. But nothing depends on it. It should not be taken to imply that this family excludes theorists that have been working in other countries.

78 This is what McKeon (1957) calls the tradition of ‘accountability’, developed mainly by Anglo-Saxon/-American philosophers.

79 The liberal notion of responsibility has traditionally been closely allied to juridical notions of responsibility. As settling the responsibility for theft or murder requires a binary choice, guilty or not, accusatory uses of responsibility have an interest in reducing ambiguities about whether someone did or did not consent to the social contract, about whether someone did or did not resist the authorities. To the extent that the liberal notion of responsibility is modelled on the juridical notion, it has, therefore, to be used in the sense of accusing or blaming anyone who is claimed to have performed some wrong. Therefore, ‘blameworthy’ is often used, and sometimes also ‘guilty’, as a synonym for ‘responsible’. It is not part of my argument to develop a theory of responsibility, but at least I want to emphasise that I do not model responsibility on the juridical notion. I think we should distinguish between matters of assessing responsibility and matters of *how* we hold people responsible. Conflating ‘blameworthiness’ and ‘responsibility’ already takes a stance on how we should people responsible, namely blame them. Again, this is not the implication that I endorse.

Rawls' view on responsible agency. Rawls does not talk much about responsibility and does not even use 'responsibility' as a central term.⁸⁰ But I do think that he presupposes a notion of responsible agency when he remarks that 'citizens think of themselves as self-originating sources of valid claims'.⁸¹ I take this remark to imply that people are basically claim-makers.⁸² Rawls realised that the world which people make claims about, and the language they use to make claims, is complex, equivocal, indeterminate, and that, consequently, people may differ in their assessments and interpretations of both facts and values. Rawls called this 'the burdens of judgement'. We cannot, therefore, in advance assume that people will agree on the claims they make.⁸³ Citizens, even reasonable ones, will arrive at different views on doctrines of the good life (comprehensive doctrines). Furthermore, that people are claim-makers does not necessarily mean that people are being respected for this. Whether people are equal is also a point of disagreement between comprehensive doctrines: 'To take an extreme case, slaves are human beings who are not counted as sources of claims, not even claims based on social duties or obligations, for slaves are not counted as capable of having duties or obligations'.⁸⁴ The question that troubles Rawls and other moral-political philosophers is, therefore, how to cope with civil disagreements within society.

The idea of people as claim-makers is key not only to understanding the pluralism of doctrines, and therefore the *disagreements* between citizens, but also to its solution: true dialogue, an ideal which many moral-political theorists that deal with the problem of pluralism are committed to.⁸⁵ It might be claimed that true dialogue is realised in the original position, as Rawls explained it.⁸⁶ In this position, people are supposed to recognise each other's capacity as claim-makers. Whatever citizens disagree about, this recognition invokes respect for each other as equal partners in a dialogue. Consequently, each has an equal say in matters of justice, not just in the sense that all participants in dialogue have an equal opportunity to say what they want to say, but also in the sense

80 Rawls – in his *A Theory of Justice* – does mention 'responsibility' and 'the principle of responsibility', for which he refers to Hart (1968).

81 Rawls (1980:543). Also see Rawls (1985:242); Rawls (1993/ 1996:32); Rawls (2001/ 2011:23–24). In (1993) and (2001), Rawls formulates it slightly differently: 'citizens [...] regard themselves as self-authenticating sources of valid claims'.

82 The idea of citizens as claim-makers presents only one of three ways in which Rawls understands citizens as free persons in the original position. Rawls introduced these three ideas, which are part of his political conception of citizenship, as a response to criticisms that the description of persons in the original position is part of a metaphysical doctrine of the person. Here I will reflect only on the idea of citizens as claim-makers.

83 Cf. Rawls 1996, lecture II, section 3. Rawls, however, does not explicitly makes a connection between the fact of pluralism and the citizens as a self-originating source of valid claims.

84 Rawls 1996:33.

85 Rawls analyses true dialogue in terms of the 'reflective equilibrium'. The parties in the original position are supposed to reach an agreement on the principles of justice. This suggests that agreement is preceded by some form of dialogue. However, Rawls's *Theory of Justice*, to my knowledge at least, does not mention the word 'dialogue'.

86 For moral-political philosophers within the liberal tradition, the original position has been, ever since Rawls's use of it, a very common strategy to understand true dialogue. Cf. Galston 1982.

that each should be truly listened to because mutual respect also implies, in terms of attitudes, a mutual responsiveness for the reasons each participant gives.

Both the case of Little Rock Nine and, more broadly, the Civil Rights Movement, of which it was part, show the difficulties of realising a public dialogue, let alone a *true* dialogue. Rawls was aware of this. Rawls is usually recognised as having devoted his analytical efforts mostly to ideal theory, which assumes that people comply with the rules given by some ideal of justice. But he also made a modest contribution to non-ideal theory by discussing the issue of civil disobedience.⁸⁷ He sees civil disobedience as ‘a public, nonviolent conscientious yet political act contrary to law usually done with the aim of bringing about a change in the law or policies of the government’ (1999a:320).⁸⁸ Intriguing about civil disobedience is that, although it implies a commitment to a deliberative democracy (to ‘public reason’, as Rawls would call it) and a trust in its functioning, at the same time it expresses the very failure of public dialogue: ‘Attempts to have the laws repealed have been ignored and legal protests and demonstrations have had no success’ (1999a:327). Rawls’s account of civil disobedience can therefore be understood as expressing his recognition of the problem of power: the asymmetry in power that exists, under non-ideal conditions, between various social groups. Citizens from some social groups have the power not to comply, or only partially, with (ideal) rules of justice (in this case: rules about equality). Even more so, they have the power – to some extent at least – to ignore legal protests and demonstrations. It demonstrates the shortcomings of views on justice that remain limited to dialogue under ideal circumstances of compliance to what an ideal dialogue requires, ignoring the problems of power.⁸⁹

While this brief discussion indicates the importance of responsible agency to liberalism as a political philosophy, it is still another issue whether or how this notion is supported by the insights of the social sciences. To the extent that liberal theorists want to start from a psychologically realistic view on responsible agency, this seems to require that they are informed by the insights of ongoing investigations in psychology.⁹⁰

87 Cf. Simmons (2010); Robeyns (2008) and Sabl (2001). For a critical discussion of Rawls’s notion of civil disobedience: see Celikates (2016). *Political Liberalism* owes its central idea of an overlapping consensus to the discussion of civil disobedience in *A Theory of Justice* because this idea was introduced within the context of this discussion ‘as a way to weaken the conditions for the reasonableness of civil disobedience in a nearly just democratic society’ (verbatim both 1985:226, note 5; and 1996:15, note 17). By the time he wrote *Political Liberalism*, he had dropped, so it seems, the subject of ‘civil disobedience’.

88 In understanding civil disobedience as a non-violent act, Rawls joined contemporaneous discussions. We should not conclude from this that Rawls has found a new way of understanding of ‘bringing about a change in the law or policies of the government’ (1999:320) because the dominant perspective for his discussion of civil disobedience remains its relation to ‘laws enacted by a legislative majority’.

89 For other criticisms on precisely this point, cf. Mouffe 2005.

90 Ever since the rise of psychology as a distinct discipline, psychologists have used different outlooks or perspectives to model their theories (see, e.g. Teo 2005). Here I will be concerned mainly with psychological research that is modelled after the example of the natural sciences.

Psychological insights may, of course, help to understand the capacities that responsible human agency requires. They may perhaps even be helpful in exploring new dimensions of responsible agency. And they even contribute to understanding the nature of unjust practices in our societies. Exemplary is the research into implicit bias. Another important example has been the Psychologists' Brief in the case *Brown v. Board of Education of Topeka*. This amicus brief was written by some 35 social scientists, including Gordon Allport, and presented a summary of the latest social-scientific insights on the negative, psychological impact of segregation on African American children.⁹¹ Whatever its exact influence on the Court, from its decision, it becomes clear that the psychological impact was included as a relevant aspect in its considerations.⁹²

But the use of psychological insights seems to have had a drawback on liberalism as a political philosophy. What is striking about the notion of responsible agency within philosophical liberalism, at least within its liberal Anglo-American tradition, is that it seems to have lost its importance to theories of justice. Scheffler, for example, observes that 'none of the most prominent contemporary versions of philosophical liberalism assigns a significant role to desert at the level of fundamental principle' (Scheffler 1992:301). Liberal philosophers do not rely on any substantial (or, as Scheffler calls it, *pre-institutional*) notion of responsible agency for their theories and even seem to be reluctant to develop one: 'it is a striking fact that the dominant contemporary philosophical defenses of liberalism, by virtue of their reliance on a purely institutional notion of desert, do indeed advocate a reduced conception of responsibility' (Scheffler 1992:314).⁹³ What explains the gradual erosion of the notion of responsible agency, as suggested by Scheffler, is that already for a longer time and at a fundamental level within philosophical liberalism itself, especially to the extent that it is embedded within Anglo-American philosophy, a Trojan horse was brought in, which is that liberal philosophers have themselves internalised, often tacitly, a *naturalist outlook* on the world by accepting the pervasive idea in contemporary (Anglo-American) philosophy that 'human thought and action may be wholly subsumable within a broadly naturalistic view of the world' (1992:309).⁹⁴ Liberal political philosophers have generally chosen to ignore the challenge of this naturalistic outlook, probably, as Scheffler suggests, because their theories of justice do not depend not on any pre-institutional notion of

91 On how psychologists became involved in this case: see Kluger (1975).

92 For a discussion of the influence of the Psychologists' Brief in this case: see Wrightsman (2008).

93 Scheffler focuses his critical discussion on liberal philosophers who, like Rawls, are not utilitarians. This is not because he favours utilitarianism. On the contrary. Their moral psychology is problematic too, he thinks: 'the ideas of reciprocity and mutuality have no fundamental ethical significance for utilitarianism, and as a result, utilitarian moral psychology seems forced to make implausibly heavy demands on the human capacity for sympathetic identification. In so doing, utilitarianism seems to underestimate the psychological constraints on the design of stable social and political institutions, and Rawls's account has great force by comparison' (1992:315-316).

94 While 'the commonly held view', as De Caro and Voltolini put it, is that 'Liberal Naturalism is not a genuine metaphilosophical option' (2010:69), in their article, they make a proposal for reconciling philosophical liberalism and naturalism. I do not think it is convincing (also see chapter 5, 7).

responsible agency, which seems to safeguard liberal political philosophers from any problems their naturalistic outlook might raise. According to Scheffler, the naturalistic outlook of liberalism nevertheless poses various problems to philosophical liberalism. Its naturalistic outlook seems to commit liberal political philosophers, for example, to determinism. This is a problem for philosophical liberalism not just at the level of theorising (for example: how to conceptualise the compatibility of determinism and agency?), but also at the moral-political level as informing and inspiring actual debates on issues of justice. What Scheffler suggests here is that liberalism's claims to neutrality towards various conceptions of the good lose credibility as long as it remains sceptical about individual agency because 'a purely naturalistic understanding of human life *is* contentious. The modern world is deeply divided in its attempt to come to terms with the power of naturalism' (1992:318).

If Scheffler's diagnosis is correct, it would suggest that philosophical liberalism allows no room for holding people responsible for their practices of hidden prejudice. My suspicion is that the liberal notion of responsible agency presupposes a mentalist framework that has become part of a naturalist programme, as indicated in the previous section. If that suspicion is correct as well, the liberal, mentalist understanding of responsible agency has the tendency not only to weaken, as indicated by Scheffler, but even to dissolve the very notion of *responsible* agency.

The problem with the notion of responsible agency and its place within liberalism as a political philosophy is illustrated by Rawls' view on responsible agency. Contrary to what Scheffler claims, it seems that Rawls does offer a notion of responsible agency, as we have seen earlier, one which views people as claim-makers. But I think this view only confirms Scheffler's suspicion that liberal philosophers like Rawls are reluctant to deal with the challenge of a naturalistic outlook. His *Political Liberalism* (1993), published one year after Scheffler's article, sheds some light on how Rawls chooses to deal with that challenge. At some point in his argument, he distinguishes between two forms of psychologies: one philosophical, the other not (the heading of the section where he brings forward his strategy is called 'Moral Psychology: Philosophical not Psychological'). He explains making this distinction as follows:

'its political conception of justice is a normative scheme of thought. Its family of fundamental ideas is not analyzable in terms of some natural basis, say the family of psychological and biological concepts, or even in terms ...' (1993/1996:87-88)

And he adds:

'political philosophy is autonomous [because] we need not explain its role and content scientifically, in terms of natural selection, for instance' (id.:88).

From a note Rawls adds to the last remark, it becomes clear that he opposes attempts, in this case by Allan Gibbard, to naturalise moral and political philosophy. Apparently, Rawls hopes to keep naturalism at a distance by introducing a distinction between philosophical and non-philosophical psychology. His theoretical framework fits only with the first. But Rawls's distinction is only a *strategic* retreat from the 'naturalistic view of the place of human beings in the world'. The effect of this strategy is that it only preserves what Hare has called the 'moral gap'. While scientific (psychological, sociological and other) research in the last two centuries has increasingly succeeded in explaining human beings in terms of natural and social forces, moral theories have not necessarily kept pace with them. This has created a distance between notions of responsibility and notions of agency. Viewed from the concerns of moral theory, this constitutes 'a gap between the moral demand on us and our natural capacities to live by it' (Hare 1996:1).⁹⁵ When Scheffler observes that 'one of the defining features of modern life is a deep uneasiness about what place there may be for ourselves and our values in the world that science is in the process of discovering' (1992:318), I take him to be recognising this moral gap. Do scientific insights attribute to people the cognitive-psychological capacities that morality requires them to have? Rawls's strategic retreat, therefore, merely raises philosophical questions on the relationship between both types of psychology. In other words: the problem of the moral gap is not yet solved.⁹⁶

While Rawls's views have immensely influenced liberalism as a political philosophy, with their focus on ideal theory and distributive justice, they have also diverted attention away from designing a more robust notion of responsible agency that is suitable to analysing issues of justice – perhaps better: injustice – under non-ideal circumstances. At this point, my argument may prove to be relevant to philosophical liberalism. If I succeed in achieving my research aims, my argument will suggest that philosophical liberalism has been mistaken in discrediting responsible agency. It may also point the way to a road not yet taken: a reevaluation of responsible agency by embracing a pragmatist framework as proposed in my argument.

95 See Hare (1996), who critically discusses various naturalistic ethics that try to substitute God, particularly those by: Donald Campbell, David Gauthier and Allan Gibbard. Also, see Smiley (1992), who also critically discusses various attempts, from a naturalist perspective, of finding a new objective moral perspective, in particular: John Harris, John Casey and Dennis Thompson.

96 Concerns about a moral gap at least motivated other philosophers, such as Strawson (1962) and Frankfurt (1971), to come to grips with the challenge of a naturalistic outlook, even if they did not analyse the problem within the context of political philosophy. For Strawson, this concern was expressed in what he called 'participant (reactive) attitudes' while for Frankfurt, it was a matter of 'our most humane concern', i.e. being a person and all that it implies: having freedom and responsibility.

1.6. Outline of chapters

My discussion in the next chapters will proceed as follows. In chapter 2, I explore the liberal notion of responsible agency and suggest how philosophical liberalism tends towards a mentalist approach in understanding this notion by conceiving (inner) deliberation as a mentalist process. In chapter 3, I introduce the notion of hidden prejudice and explore both a mentalist and pragmatist (in this case, Bourdieusian) version of it. At this point of my argument, I will still assume that both versions are compatible. Here I will conclude that the core problem with practices of hidden prejudices is not so much that they have a discriminatory impact, but they are relatively insensitive to social change because they also include practices of silencing. In chapter 4, I will explore the assumptions underlying a mentalist approach towards both responsible agency and the veil of prejudice. In chapter 5, I will criticise the mentalist notions of rationality and freedom, use this criticism for rejecting the mentalist framework and suggest an alternative framework: pragmatism (now in the Brandomian version of it). In chapter 6, I will use this pragmatist framework to propose a different pragmatist framework of the veil of prejudice. Finally, in chapter 7, I will also use it to propose a pragmatist notion of responsible agency and, more specifically, epistemic autonomy, from the perspective of both the person whose epistemic autonomy may be at stake and the person who, in performing a non-deliberative action, might violate (or not) the first person's epistemic autonomy. This chapter will be concluded with a discussion of two case studies.⁹⁷

97 Chapters 2 to 7 are to some extent symmetrically organised, with chapters 4 and 5 functioning as the hinges between chapters 2 and 3 on the one hand and chapters 6 and 7 on the other hand. Chapters 2 and 7 explore the notion of responsible agency, chapters 3 and 6 exploring the 'veil of prejudice' and chapters 4 and 5 exploring notions of freedom and rationality. The first chapters (2 to 5) are primarily dedicated to the mentalist framework, while the later chapters (5 to 7) are mainly concerned with the pragmatist framework.

2

Chapter 2

The liberal notion
of responsible agency

2.1. Introduction

This chapter aims to introduce *autonomy* as the standard liberal notion of responsible agency and to trace what might constitute its epistemic dimension, i.e. to delineate within this notion of autonomy a more specific notion of *epistemic* autonomy. First, I introduce two basic features of the liberal notion of responsibility: *authorising*, which represents the *epistemic* dimension of liberal responsibility, and *reflection*, which represents its *psychological* dimension. To illustrate how autonomy might work by bringing together ‘authorisation’ and ‘reflection’, I will then discuss one specific, influential account of autonomy, one that was proposed by Harry Frankfurt (1971). I will present it as the notion of *reflective volitional autonomy*, one that understands *authorising* as a reflective, specifically hierarchical, process (section 2). Then I will introduce the problem of manipulation (section 3). This appeals mainly to the psychological structure of the previous notion of autonomy. Introspective self-knowledge may be understood as a response to this problem, conceptually at least. Here I will present the notion of *reflective epistemic autonomy* (section 4). Finally, I will argue that reflective epistemic autonomy, as a necessary condition for this standard liberal notion of agency, should be understood as embedded in a mentalist framework and, therefore, as *introspective* epistemic autonomy (section 5).

2.2. Reflective volitional autonomy

Liberalism can be viewed as a successor to early modern contractarianism. Liberal theorists are, therefore, in a sense neo-contractarians.¹ Early modern contractarians had introduced the idea of individual freedom as responsibility,² which they thought of in practical terms: as promising or signing a contract (even if this was only metaphorically).³ Introducing this idea helped them rethink society’s political-theoretical basis, which they reimaged as a social contract based on each individual’s decision (‘consent’, ‘agreement’) to authorise the social contract.⁴ But the idea of authorising as promising raised its own problems.⁵ The various roads that philosophers explored to solve them

1 Here I follow Santoro (2003).

2 Today, the liberal notion of responsibility often goes by the name ‘moral responsibility’, mostly in debates on free will and determinism, or ‘desert’ (sometimes ‘individual desert’ or ‘personal desert’), in debates on political and legal theory. For desert: see Sher (1987); Pojman and McLeod (1999). For moral responsibility: Fischer and Ravizza (1993). In chapter 3, more specific references will follow. A large part of these debates is devoted to the issue of agency, as a fundamental requirement for (moral) responsibility as their dogma is that responsibility presupposes agency: without agency, no responsibility. Sometimes theorists are so much absorbed by the problem of agency that they lose sight that freedom is part of the broader notion of responsibility, a claim to which I will return in chapter 5.

3 As such, it obviously has its roots in juridical thinking. See, e.g. Black (1993).

4 See, e.g. Baumgold (2010).

5 For a discussion of such problems, specifically in relation to Hobbes’ social contract theory: see Hampton (1986).

resulted in the first contours for what would become a liberal notion of responsible agency as *autonomy*.⁶ However, both the early contractarian and the liberal notions of responsible agency turn on the idea that the individual is the primary authority, expressed in each individual's 'authorising' the social contract. This is the crucial and founding insight of early contractarianism and liberalism.

In this chapter, I will hone in on 'autonomy' because this has become a central concept of philosophical liberalism.⁷ It has acquired this prominent role since the early 1970s.⁸ But it has been given many different meanings, as several writers have observed.⁹ Some uses of the term 'autonomy', for example, resemble what Berlin calls negative freedom.¹⁰ But it usually comes more closely to what Berlin meant by 'positive freedom'. In this chapter, I will therefore use the terms 'autonomy' and 'positive freedom' interchangeably.¹¹

6 Scholarship in recent decades has acknowledged that early modern debates on morality were premised on the need for an adequate, moral notion of subjective freedom or self-governance. See e.g. Darwall (1995); Beiser (1996); Schneewind (1998); Gill (2006). Whereas Darwall, Beiser and Gill focus on British moralists, Schneewind tried to show how 'self-governance' was central to a European scene of theorists. For example, Korsgaard (1996a) follows this line too.

7 See Milligan and Miller 1992; Santoro 2003; Christman and Anderson 2005; Colburn 2010; Sneddon 2013 (chapter 7).

8 For historical overviews of the modern use of autonomy: Hinchman (1996); Santoro (2003); Schneewind (2013). For historical overviews tracing the (late medieval and) early modern origins of autonomy: Charles Taylor (1989); Hinchman (1990); Schneewind (1998); Santoro (2003). While such overviews tend to present a 'relatively unified intellectual history of subjectivity', Yousef (2004) instead offers a reinterpretation of some philosophical and literary classics on autonomy as showing its limitations, rather than promoting it. For conceptual overviews of autonomy (but also providing some historical insight): Dworkin (1988); Christman (1988); Christman (1989); James Taylor (2005); Christman and Anderson (2005); Christman (2015).

9 For attempts at either giving an overview or delineating a central meaning of autonomy: see Feinberg (1986, chapter 18, 'Autonomy'), Faden and Beauchamp (1986), Dworkin's (1988), Johnston (1994, chapter three, 'Perfectionist Liberalism'), Arpaly's (2003, chapter 4, 'Varieties of Autonomy'). Especially Dworkin (1998), Arpaly (2003), but also O'Neill (2003) have expressed doubts about whether there is an agreement on the notion of autonomy. In the main text, I explain how I understand 'autonomy'.

10 For the use of the term 'positive liberty': see Crocker (1980), who argues for a notion of freedom that is, according to Kristjánsson (1996), a notion of *negative* freedom. For the use of the term 'autonomy': see Sunstein (2014), who explicitly uses the term 'autonomy' but understands it as 'freedom of choice' and, on closer reading, as non-intervention. The sources for Sunstein's use of the term trace the notion of autonomy back to Joel Feinberg (Conly 2013) and J.S. Mill (Wright and Ginsburg 2012). Conly's reading of Feinberg's notion of autonomy confirms my suspicion that Sunstein's notion corresponds with Berlin's negative freedom (although Conly does not refer to Berlin's distinction). (I should add that Christman (2015:5) calls Feinberg's notion 'basic autonomy', although he is clear it should be kept separate from 'personal autonomy'). Wright and Ginsburg's reading of Mill is more ambiguous (they refer to Amartya Sen's 'decisional autonomy', which is his term, to some extent, for Berlin's positive freedom). But their remark on Sunstein's notion (2012:1069, note 159) confirms my suspicion again.

11 Here I follow Christman (1989:3 'that concept's [of positive freedom] identical twin, individual autonomy'). Also, see Santoro, who, a decade later and borrowing Christman's phrase, again observes that '[m]any authors have taken Berlin's notion of 'positive liberty' to be the 'identical twin' of the notion of autonomy' (2003:1). Dissatisfaction with liberal autonomy has induced several theorists to develop different kinds of notions of autonomy which, in my view, come nearer to what Berlin calls a negative notion of freedom. Among these would be, my suggestion would be, the account of autonomy as developed by Meyers (2005) and perhaps too the notion of 'political autonomy' that Rawls introduced in his *Political Liberalism*. However, it is beyond the scope of my present purposes to provide a detailed argument for these claims.

Liberal autonomy is usually linked to some ideal of individuality.¹² Its various versions are sometimes referred to as Real Self Views.¹³ They share the idea that autonomy consists of *self-realisation* (self-determination; self-direction): ‘the capacity to be one’s own person, to live one’s life according to reasons and motives that are taken as one’s own’ (Christman 2015). They are often referred to as ‘personal autonomy’, ‘individual autonomy’ or simply ‘autonomy’.¹⁴ For the purpose of my argument, I understand liberal autonomy in a more extended sense. It also includes proposals for autonomy that various Anglo-American philosophers modelled or based on Kant’s notion of autonomy. And I think it should even include Rawls’s notion of ‘political autonomy’.¹⁵ The term ‘liberal autonomy’ will be used to refer to all these different versions.¹⁶ Their common assumption is that people have the capacity for self-reflection. They also share the idea that people’s agency (or: freedom) is the outcome of a certain qualified process of reflection, even if they differ on what counts as a qualified process.

During the modern period, many philosophers have worked on notions of autonomy, although not always within the context of political philosophy.¹⁷ The main innovation of these notions, as compared to the early contractarian notion of negative freedom, was the introduction of the idea that people’s ‘inner life’ (subjectivity) has a hierarchical structure: one part is identified with their ‘higher nature’, another part with

-
- 12 See Hinchman (1990), who traces the ancestry of this type of autonomy via Mill and Humboldt back to Goethe, Schiller and the early German romantic movement.
- 13 For this term: see Wolf (1990, esp. chapter 2).
- 14 Notions of personal autonomy are sometimes considered different from notions of ‘moral autonomy’ when this last category is understood in a Kantian sense. See Waldron (2005) or Flikschuh (2013). By itself, the phrase ‘moral autonomy’ does not, however, unambiguously refer to either personal autonomy or Kantian autonomy. See the definition by Dworkin: ‘The most general formulation of moral autonomy is: A person is morally autonomous if and only if his moral principles are his own’ (Dworkin 1988:34). This comes much closer to notions of personal autonomy than to those of moral (Kantian) autonomy as defined by Waldron (2005:307): ‘A person is autonomous in a moral sense when he is not guided just by his own conception of happiness, but by a universalized concern for the ends of all rational persons’. The problem suggested by this definition is what I describe, in the next section, as the problem of how liberalism understands the moral structure of human interactions.
- 15 In his *Political Liberalism* (1993), Rawls distinguishes between *political autonomy* and *moral autonomy*. (The second is Rawls’s term for what is often referred to as individual or personal autonomy.) As he explains in his introduction to *Political Liberalism* (in the 1996 edition), this ‘distinction is unknown to *Theory*, in which autonomy is interpreted as moral autonomy in its Kantian form, drawing on Kant’s comprehensive liberal doctrine’ (1996:xlili, note 8). Rawls introduces the distinction to criticise his own use of autonomy in *A Theory of Justice* and replace it, in *Political Liberalism*, with a different notion: *political autonomy*. For a criticism of his shift from moral to political autonomy: see Abbey and Spinner-Halev (2013).
- 16 I suspect that what here I call ‘liberal autonomy’ agrees with what Zheng calls *responsibility as attributability* (2016:63-65, section 2).
- 17 In referring to pre-Kantian concepts of autonomy, Schneewind (1998) uses the term ‘self-governance’.

their ‘lower’ nature.¹⁸ As their names already suggest, the relation between both elements is understood in a hierarchical sense: the higher nature is supposed to have a higher value than the lower nature because the higher nature, as Berlin puts it, ‘aims at what will satisfy it in the long run, with my ‘real’, or ‘ideal’, or ‘autonomous’ self, or with my self ‘at its best’” (2002:179). Underlying the notion of autonomy is the idea that human subjectivity is structured by the hierarchical relationship between two different ‘selves’. In this sense, Berlin’s analysis is not controversial. In fact, perhaps partly due to Berlin’s analysis itself, it has become the standard view.¹⁹ Hinchman (1996:490), for example, although he uses somewhat different terminology, observes in theories of autonomy the same ‘basic pattern’ as Berlin does:

‘There is first a ruling, law-giving, controlling, or evaluating element (the *nomos*) situated within the individual (for, if it were outside, autonomy would not even get off the ground). The controlling or evaluating aspect of the self must relate itself to another internal element that, by definition, would vitiate personal autonomy if it were not directed, ruled, or evaluated critically.’

Various scholars have commented that this model is indebted to the older, Stoic model.²⁰

The idea of subjectivity as having an inner hierarchical nature served as the conceptual background for working out solutions to various problems. One of them was the problem of how we should understand ‘consent’ (in its pragmatic sense) as backed up by an appropriate, *subjective* sense of authorising.²¹ But there was still another variety of problems that together originated in questions that early contractarian views had raised

18 Various scholars have acknowledged this structure (mostly using the term ‘hierarchical’): see Thalberg (1978); Friedman (1986); Christman (1988); Bratman (2003). Santoro (2003) uses the term ‘hierarchical-dualist’. He uses it as a generic term. For those notions of autonomy influenced by Frankfurt’s account (1971), he even uses the more specific term ‘double-order of desires’. Santoro claims that the hierarchical-dualist anthropological notion is a liberal one (of which he is critical) and traces it back to Locke. Hinchman (1990) claims it entered Anglo-American philosophy due to Humboldt’s notion of autonomy, in Mill’s reception, and that in his turn, ‘[f]rom Kant and Fichte, Humboldt appropriated a thoroughgoing metaphysical dualism that described human nature in a series of dichotomous relationships: reason and sense, moral obligation and inclination, ideal man and empirical man, abstract understanding and sensibility, and freedom and necessity’ (1990:762–763). The phrase ‘homo duplex’, originally invented by Durkheim in a different context, is sometimes used as well to refer to hierarchical-dualist subjectivity, for example, by Hunter (2001) as part of his criticism of the Kantian metaphysical-moral philosophy in favour of a less well-known tradition of civil philosophy which he traces to Pufendorf and Thomasius. Both Hinchman (1990) and Cuypers (2000) point out that modern notions of autonomy are rooted in two different traditions: the Enlightenment tradition of rationality and the Romanticism tradition of authenticity.

19 See e.g.: Charles Taylor (1989); Christman (1989); Hinchman (1996); Schneewind (1998); Santoro (2003); James Taylor (2005).

20 For the link between autonomy and Stoicism: see Cooper (2003).

21 See, e.g. Schneewind’s *The Invention of Autonomy* (1998) which is an attempt to reconstruct various pre-Kantian notions of autonomy in terms of precisely solving this problem. For some critical notes on Schneewind’s reconstruction: see Swaine (2016).

about human nature, i.e. the ‘empirical self’ (a phrase Berlin uses). One such problem was about *acrasia*, i.e. weakness of will. In Santoro’s interpretation, this problem turned into a political one for the social contract since Locke: ‘Locke’s novelty is that *acrasia*, the ‘weakness of the will’, is no longer as in Augustine an existential problem concerning individual persons only: it becomes a social problem, indeed *the* problem of political order’ (2003:133).²² Another problem was what Schneewind (1998) called the Grotian problem: what created for Grotius the problem of social order and stability was that people are both self-interested and sociable. It provoked much discussion during the early modern period, as Schneewind shows, about how human nature should be really interpreted. Was the Grotian problematic correct? Were human beings self-interested by nature, but were the effect of their self-interested actions social in nature (Mandeville)? Were human beings corrupted by civilisation (Rousseau)? Many different analyses of human nature were suggested. One of the threads running through them was how people could keep their unsociable feelings, their egoistic emotions, under control. In this sense, Kant rephrased the problem as one of people’s ‘unsocial sociability’ (*ungesellige Geselligkeit*).²³

An important key to solving these problems about the proper understanding of human nature was, and still is, the notion of ‘control’ (or perhaps: ‘self-control’). Although I do not intend to discuss these problems, I do want to make some comments on this notion of control, as it will play an important role in the background of various debates that are yet to be discussed: one about freedom of the will (this section), another about manipulation (in the next section) and yet another about freedom as opposed to determinism (next chapter).

Both *authorising* and *control* play a role in the design of various proposals for autonomy, but they should not be conflated. The first notion is common to both early contractarian and liberal notions of freedom, even if authorising, in the case of early contractarian freedom, is assimilated into ‘consent’. However, the second notion, *control*, can only be theorised if we assume a hierarchical view of people’s subjectivity. Adherents of personal autonomy have identified authority with the *real self* (or: *ideal self*): one’s self provides the source for authorising one’s actions. But authority is not the same as control. If a person has found out her ‘ideal self’ and now wants to lead the life that is true to this ‘ideal life’, she will still have to master any distressing emotions that might distract her from achieving her goal: realising her ‘ideal self’. She will have to learn to control these emotions. This problem is reminiscent of what troubled ancient Stoics. The solution to the problem turns on mastering your self. While authorising (self-realisation) is not possible without control (self-mastery), the reverse is also not possible: self-mastery without self-realisation too leads to nothing.

22 Interestingly Santoro tries to reconstruct various pre-Kantian notions of autonomy in terms of solving the problem of *acrasia*.

23 See Kant (1784b/ 1964:37). Also, see Wood (1991).

While the notion of control is, in my view, a feature of positive freedom, Skinner suggests that it is a feature of an extended notion of negative freedom rather than positive freedom. One of the phrases Berlin uses to pick out what is distinctive about notions of positive freedom is 'self-mastery'. Skinner takes him to be interested really in 'mastering your self' in the sense that 'the obstacles to your capacity to act freely may be internal rather than external, and that you will need to free yourself from these psychological constraints if you are to act autonomously' (Skinner 2002:239). I refer to this as the notion of control (Skinner does not use the term 'control'). As 'mastering your self' turns out to be a mere internal, psychological version of the external version of negative freedom, we are 'still speaking in consequence about the idea of negative liberty' (Skinner 2002:239). Therefore, Skinner claims, against Berlin, that the idea of 'mastering your self' fails as a qualification of positive freedom.

I think Skinner is mistaken, in part because he identifies positive freedom entirely with self-realisation (*authorising*) while ignoring that it admits of other features, but more importantly because he misrecognises what is really at stake in Berlin's argument against positive freedom. Skinner transfers the idea of 'non-interference', the criterion Berlin uses for defining negative freedom, to the realm of subjectivity, the inner world of people. In transferring this criterion, he uses it in a metaphorical sense. This may seem an innocent move, but it conceals a deeper-lying shift in the criteria for assessing freedom. While the *notion* or the *term* 'non-interference' can be used to describe the assessment of *freedom from* internal obstacles, the actual criterion for this assessment is '*control*'. This becomes clear once we try to describe the assessment of freedom as non-interference, the other way around, in terms of control. Freedom as control over other people's actions is not the same as freedom as non-interference. The important difference between both forms of freedom is that freedom as control depends, while freedom as non-interference does not, on the assumption that subjectivity is hierarchical-dualist. This is the principal reason for Berlin to reject a notion of freedom that relies on control. Berlin is not interested in freedom as what expresses a 'will' that is free from psychological obstacles. For Berlin, freedom is exactly the freedom also to do what is irrational, or stupid, or wrong.²⁴

Kant is often mentioned, beside J.S. Mill and to a lesser extent Humboldt, as a philosopher whose writings are an important source for philosophical liberalism: 'Kant's moral principle that individuals cannot be sacrificed or used without their consent to achieve other goals becomes a political principle' (Santoro 2003:13).²⁵ Kant introduced

24 I reverse here the claim he ascribes to adherents of positive freedom: 'Freedom is not freedom to do what is irrational, or stupid, or wrong'.

25 For Kant: see e.g. Wolff (1976), Nozick (1974), Gewirth (1978), Lindley (1986). For a general discussion of moral autonomy: see Dworkin (1988, esp. chapter 3). In an overview of what 'English-language moral philosophers have had to say about autonomy after Kant' Schneewind mentions, besides Mill, also Emerson as an early philosopher who argued for 'self-reliance', a notion that also belongs to the family of notions on moral autonomy. Dworkin does not mention Emerson, but Emerson's notion of self-reliance still appeals to some modern philosophers. See, e.g. Kateb (1995); Button (2014). Mill is mentioned, beside Kant, by Rawls (in the introduction of the (2005) edition of *Political Liberalism*). Mill, in his turn, had read Humboldt.

the term ‘autonomy’, which has now become a technical term, and used it for his own notion of autonomy as self-legislation. But references to his notion of autonomy can be misleading because liberal autonomy differs from Kant’s notion in important respects.²⁶ First of all, what is ascribed to Kant as his notion of autonomy, may sometimes be ambiguous as Kant argued for self-governance in various, different ways.²⁷ Whereas for Kant, autonomy expresses his concern with making explicit the moral connection between all human beings, for liberal theorists, the notion of autonomy expresses their deepest concern with the integrity of the individual against the possible intrusions, interferences and dominations of others. We may characterise Kant’s autonomy as representing a morality of universal personhood as the basis for human dignity. Instead, liberal autonomy represents a morality of individual personhood (or: individualism): individuals are the prime holders of rights in society and the prime sources of moral values.²⁸ But there is something special about this individualism that did not obtain for early modern social contract theorists: what is central in this individualism is an ideal of individuality: ‘What a person is, what he wants, the determination of his life plan, of his concept of the good, are the most intimate expressions of self-determination, and by asserting a person’s responsibility for the results of this self-determination, we give substance to the concept of liberty’ (Fried 1978:146).²⁹

The references to Kant’s notion obscure a deeper problem, I think. Liberal philosophers like Rawls sometimes place themselves explicitly within the philosophical tradition of Kant’s ethics. But in contrast to Kant, they assume that people, in what characterises their personhood, are moral subjects that are essentially not noumenal but empirical (in Kant’s vocabulary: phenomenal).³⁰ This may be no surprise. Liberalist theorists usually have taken an interest in the development and progress of the sciences, particularly of the natural sciences. They believe that philosophical theory should take this into account, also in their concepts of subjectivity.³¹ Understanding people as

26 O’Neill (2003:3) suggests that the ‘principal source for most conceptions of *rational autonomy* is, I think, not Kant, but John Stuart Mill’s *On Liberty*’.

27 See Larmore (2013). One source he mentions for one of Kant’s views on self-governance is Kant’s essay: ‘Beantwortung der Frage: Was ist Aufklärung?’ (1984). Interestingly I could find in the literature on autonomy scarcely any reference to this essay. For example, references are lacking in these general discussions on autonomy: Christman (2015); Christman (1989); Christman and Anderson (2005); Colburn (2010); Sneddon (2013). References are also missing in these discussions on specifically Kant’s notion of autonomy: Herman (1993); Hill (1991); Korsgaard (1996c). Exceptions are: Hinchman (1996); Santoro (2003); LaVaque-Manty (2006); Schneewind (2013). Schneewind (2013) claims about the notion of autonomy implicit ‘Beantwortung der Frage: Was ist Aufklärung?’: ‘its constructive thrust is grounded in his view of autonomy [as self-legislation]’ (2013:149). Others disagree: see Hinchman (1996); LaVaque-Manty (2006); Larmore (2013).

28 Cf. Neumann (2000).

29 Also quoted in Dworkin et al. (1997).

30 See Neumann (2000).

31 See Scheffler (1992).

noumenal subjects was in that respect not attractive: it has the ring of metaphysics.³² As a consequence, they needed to find a notion of autonomy that met the requirement of ‘empirical possibility’ (Dworkin 1988:7). For this reason, the analyses that Frankfurt (1971) and Dworkin (1976, 1981) developed at an early stage in the debate on autonomy were very appealing to many theorists because they seemed to make no metaphysical assumptions.³³ As their analyses of autonomy are closely related, they are often discussed as one and are sometimes called the Dworkin/ Frankfurt (DF) model.³⁴

Below I will discuss Frankfurt’s account (1971) in detail. It has had considerable influence on notions of autonomy.³⁵ In later chapters, I will refer to this account. Although Frankfurt does not use in his (1971) a variety of terms that I used so far, including the term ‘autonomy’, his analysis is usually taken to assume that people’s subjectivity is hierarchical-dualist. It also does not mention ‘authorising’. When Frankfurt talks about ‘identifying’, I take this to be roughly equivalent with what I have previously called ‘authorising’. The source of authority for ‘identifying’ is the second-order desire, a crucial notion to Frankfurt’s analysis, as we will see. I understand his notion of autonomy to include the two features of ‘authorising’ and ‘control’. But his analysis is mainly concerned with analysing the problem of control. The notion of ‘authorising’ (for Frankfurt: identification) is in *this* analysis assumed to be unproblematic.

Frankfurt’s account of freedom may be summarised, in his own words, as follows: ‘A person’s will is free only if he is free to have the will he wants’ (1971:18). His argument proceeds in two stages: first, Frankfurt explains the difference between which sort of being is a person and which is not. He then goes on to make clear when a person is free and when she is not.

Frankfurt uses the examples of two narcotics addicts. Before he explains why one

32 On the influence of Kant on liberal political philosophers: see Flikschuh (2000). Also, see Williams (1962) and Carter (2011), who both reject the idea of a ‘transcendental’ (Williams), ‘noumenal’ (Carter) self as a basis for understanding personhood. On the influence of Kant on moral philosophers: see Flikschuh’s comment on what she perceives as the dominant interpretation in Anglo-American readings of Kant (2000:50–54, esp. note 2). Neo-Kantians who are engaged in the debate on transcendental freedom are usually committed to a non-metaphysical interpretation of people as persons and to the notion of freedom as free will, which makes them Kantian compatibilists: see, e.g. Allison (1990), Korsgaard (1996c), Bok (1998), Nelkin (2011). It is beyond my present scope to discuss the problems with their readings of Kant but see Schönecker (2013). Exceptions among Anglo-American Neo-Kantians are Karl Ameriks and Eric Watkins: Blöser (2014:237–240).

33 This also seems to hold for Anglo-American, *Kantian* philosophers. See, e.g., Korsgaard (1996a), who acknowledges her debt to Frankfurt also, at some point in her discussion, avows that ‘[i]n one sense, the account of obligation I have given [...] is naturalistic. It grounds normativity in certain natural – that is, psychological and biological – facts’ (1996:160).

34 Christman introduced the phrase ‘DF model’ (1988:112). Neely (1974) is also mentioned by James Taylor (2005:1) as an interesting and one of the earliest contributions to the contemporary debate on autonomy, but it has been less influential.

35 See Christman (1989); Santoro (2003); Schneewind (2013). On the influence of Frankfurt in a broad range of debates: also see this chapter, section 6. Weinstein (2007) claims that Frankfurt’s account of free will resembles T.H. Green’s account in his (1886). For his discussion of the relevance of these similarities for current debates on autonomy: see Weinstein (2007:101–107).

is a person and the other is not, he describes in what respect they are *no different*. He assumes, for example, that ‘the physiological condition accounting for the addiction is the same in both men, and that both succumb inevitably to their periodic desires for the drug to which they are addicted’ (1971:12). Their desire for the drug belongs to what Frankfurt calls ‘first-order desires’ which represent a certain type of mental states: the *conative attitudes*, i.e. the wishes, wants, volitions, or whatever we like to call them, that induce certain beings to perform, or even abstain from, a certain action. Frankfurt does not mean to say that the desire for the drug (the ‘first-order desire’) in the examples of his narcotics addicts necessarily leads to their drug addiction. He considers that beings (people or animals) may have various first-order desires at the same time. A conflict of these desires may delay, or even prevent, the translation of one specific desire into action. The possibility of a conflict of first-order desires is also a common feature of the two addicts.

Whatever other common features the two narcotics addicts may have, they are not the same kind of creatures: one is a person, and the other is not. What is special about persons is ‘the structure of a person’s will’ (1971:6). A person may have second-order desires. She has one if she wants a certain first-order desire. Having second-order desires as such is not yet crucial to being a person. As a special subcategory of the second-order desires, Frankfurt recognises second-order *volitions* that a person has if she wants a desire *to be her will*. Only having second-order volitions is constitutive of being a person.³⁶

Once we turn again to the examples of the addicts, it may seem puzzling what difference the second-order volitions would make. Both addicts may experience a conflict of desires. In that case, the first-order desire for the drug will settle the conflict for both of them. After all, they both are addicts. In this respect having second-order volitions makes no difference. And yet, there is a difference to having a second-order desire. We need to have a better understanding of second-order volitions. The addict who is not a person is indifferent about any of his desires. Therefore, he is indifferent about what ultimately moves his will to act. Frankfurt famously called this a ‘wanton’: someone who cannot or does not care about his will to act. On the other hand, the addict who is a person:

‘identifies himself [...] through the formation of a second-order volition, with one rather than with the other of his conflicting first-order desires. He makes one of them more truly his own and, in so doing, he withdraws himself from the other.’ (1971:13)

All his first-order desires are his, in some sense. In this sense, the desire to take the drug is also his. When the desire to take the drug takes over and determines the will to act, the addict is therefore not moved by some external influence. But as he identifies

36 Frankfurt allows for the possibility that a wanton may have second-order desires but no second-order volitions (1971:11, above).

himself with one desire in particular: the desire *not* to take the drug, he makes this desire *his*, but in a different sense, and he withdraws himself from that other desire: the desire to take the drug, almost as if this were, again in this different sense, *not his*. In Frankfurt's example, it does not help him overcome his desire to take the drug. He is an unwilling addict. While his identification does not make a difference for the outcome, and neither perhaps for anyone observing him, it does make a difference to himself because he cares for his identification, for his will and for the actual outcome. And that is distinctive for a person: not just having the right structure of will, one that includes second-order volitions, but having the right structure of will that *matters to herself*. It brings out another dimension of personhood: the first-person perspective as one that is distinctive for persons. This point remains only implicit in Frankfurt's own account, but it has been accentuated in, for example, Baker's reading of Frankfurt: 'To be able to think of oneself as oneself is not just to have a perspective or subjective point of view — dogs have perspectives — but also to be able to think of one's perspective as one's own, and to think of others as having different subjective points of view from one's own' (1998:331). Identification and withdrawal, first-person perspective: whatever we call it, these are different ways of claiming that *ownership* is a necessary condition for freedom, one that is constituted by the reflective structure of the person's will.

Being a person, i.e. having second-order volitions, is not what by itself counts as having freedom of the will. Frankfurt chooses to understand freedom of the will as analogous to freedom of action. Someone has freedom of action if he is free to do what he wants to do. Freedom of the will is explained accordingly: someone has this freedom if he is 'free to want what he wants to want' (1971:15). What matters for both forms of freedom is not the coincidence between want and action or between second-volition and first-order desire. This may be 'not his own doing but only a happy chance'. Therefore, what matters is *effectiveness*. A creature has free action if he succeeds in translating his want into actions. And a person has free will if he succeeds in translating his second-order volitions into his will or, in Frankfurt's own words, 'in securing the conformity of his will to his second-order volitions' (1971:15). Only then will she have the will of her own.

All other creatures, those who are not free, cannot simply be fit into one and the same category. A wanton may have freedom of action, but he cannot ever have freedom of the will. While a person may have freedom of action as well, she may also have freedom of the will. She has the capacity of enjoying or lacking a will that is her own, whereas a wanton cannot lack a will of his own because he has no second-order volitions and therefore has not even the capacity of enjoying a will of his own. Being a person, however, is no guarantee of having a free will. It leaves the possibility that a creature is a person and yet is not free because she lacks a will that is her own. One example is our unwilling addict. From this perspective, having second-order volitions or not having them seems to imply no difference between the unwilling addict and the wanton. The result is the same: they both have no free will. The difference between

wantons and persons can be explained in yet other terms: those of caring. A wanton ‘does not care about his will’ (1971:11), whereas a person does. From this perspective, it does not matter whether a person is effective or not in making her will conform to her second-order volitions. While the unwilling addict has no free will, he does care about his will. Of course, free and unfree persons differ in how they experience their human condition. Whereas the free person enjoys the satisfaction of her second-order desires, the unfree person experiences a tragic fate because ‘he finds himself a helpless or a passive bystander to the forces that move him’ (1971:17). It may therefore be claimed of a person who is unsuccessful, such as the unwilling addict, that ‘he is estranged from himself’ (1971:17), an experience that originates in the frustrations of second-order volitions. The unfree person cares about what will is hers but is alienated from her actual will, which is not her own. This condition of estrangement does not exist for wantons because, once again, they have no first-person perspective in the relevant sense: for the wanton, ‘the freedom of his will cannot be a problem’ (1971:15).

2.3. The problem of manipulation

In the years just before the First World War, a young woman, Miss Natalija, studied philosophy at a university in Germany. At some time, a professor began to court her. She, however, was not interested in him and refused his courtship. As she would soon find out, the professor did not resign in his rejection. First, he tried to influence her in accepting him as a lover by trying to establish a friendship between her and his sister-in-law. As this failed, he took recourse to a more radical measure of influencing her. He started to use a specific sort of *influencing machine* even though the police had prohibited its use. It was made in Berlin and had the form of a human body, even more so, the form of the young woman’s own body. The inner parts of the electrical machine represent the internal parts of her body. With the machine, the professor was able to influence her. She did not know how the machine was handled, but she surmised it connected to her through telepathy.

The woman in this story is not fictional. What happened to her, being subjected to an influencing machine, was fictional neither, not in the ordinary sense at least. She really believed she had been subjected to the influences of an ingeniously crafted machine used by her professor of philosophy. She suffered from schizophrenia and had turned for help to a psychoanalyst, Victor Tausk.³⁷ Even if we can discredit her story as

³⁷ Tausk analysed her case in his article from (1919), which is devoted to analysing, from a psychoanalytic perspective, the schizophrenic delusions of various patients that centre on the existence of some machine. Usually, this machine ‘produces, as well as removes, thoughts and feelings by means of waves or rays or mysterious forces’ ((1919/ 1992:186). The delusions have a link, as Tausk suggests, with ‘the progressive popularization of the sciences’. Tausk’s examples do not present the first cases of delusions about ‘influencing machines’. The first reported case of a patient with such delusions was James Tilly Matthews (Jay 2012).

being a schizophrenic fantasy, it nevertheless illustrates the ultimate nightmare for liberal philosophers who adhere to the idea of autonomy: being manipulated.³⁸ In whatever ways they try to clarify what they mean by ‘being autonomous’, often it is *also* defined in opposition to ‘being manipulated’. In their overviews of the contemporary debate on (individual) autonomy, various authors have made this observation, for example, Hinchman (1996:488): ‘the main requirement [of autonomy] is that the choices be truly one’s own, that one must not have been manipulated, gulled, brainwashed, or conditioned into making them’; and a few years later Santoro (2003:13): ‘the concept of ‘autonomy’ normally refers to an individual who is not manipulated and is characterised by a will of her own, a capacity for pursuing her own chosen ends; ends not determined by others’.³⁹ In a more recent overview, manipulation is again included as a negative point of reference when Christman (2015:1) summarises the general notion of autonomy as ‘the capacity to be one’s own person, to live one’s life according to reasons and motives that are taken as one’s own and not the product of manipulative or distorting external forces’.⁴⁰ In a nutshell: being autonomous implies *not* being manipulated, while being manipulated implies *not* being autonomous.⁴¹

In this section, my aim is to get a clearer view of *what problem* of manipulation, i.e. what *type* of manipulation, is distinctive for the debate on autonomy (positive freedom) rather than for a debate on negative freedom or perhaps other kinds of freedom. This is no easy task. The possibility of people being manipulated has captured the attention of theorists on autonomous agency for at least five decades – and, as we will see in the next chapter, of theorists on free will agency (in the context of determinism) for even a century. And yet, theorists on autonomy are often not clear on what they mean by manipulation. Usually, they do not provide conceptual analyses of manipulation, although sometimes they do provide examples. Common examples are: hypnosis⁴², brainwashing⁴³, subliminal suggestion (as in cases of subliminal advertising)⁴⁴ and

38 Perhaps another nightmare they might fear is that they are subjected themselves to an influencing machine used, let’s say, by their students.

39 The original Italian version of Santoro’s book was published in 1999. The English version adds no literature that is more recent.

40 Also, see Taylor (2005:23): ‘it is generally accepted that a person’s autonomy can be undermined if she is successfully manipulated or deceived by another’. Also, see earlier overviews from the 1980s: Dworkin (1988); Christman (1988); Christman (1989). That ‘not being manipulated’ as a condition of autonomy has found its way into political philosophy, can be seen for example in the case of Rawls: while his *A Theory of Justice* (1971) did not contain any references to such a condition, in his *Political Liberalism* (1993), he explains at some point in his argument that ‘free and equal citizens’ are to be viewed ‘not as dominated or manipulated’ (1993/ 1996: 446).

41 This does not imply, however, that ‘not being manipulated’ is enough to count as ‘being autonomous’ because ‘to be autonomous one must also be able to make laws for oneself’ (Santoro 2003:15).

42 See e.g. Benn and Weinstein (1971); Benn (1976); Young (1980); Christman (1988); Thalberg (1978); Feinberg (1986); Taylor (2005b).

43 See Benn and Weinstein (1971); Benn (1976); Thalberg (1978); Thalberg (1979); Berofsky (1995).

44 See e.g. Benn (1967); Dworkin (1976); Thalberg (1978); Arrington (1982); Dworkin (1988); Christman (1991); Berofsky (1995); Mele (1995); Colburn (2010).

varieties of engineering the brain (brain surgery, implanting all kinds of devices)⁴⁵. Even a distinct philosophical debate on manipulation itself is relatively new, starting only in the 1980s.⁴⁶ This debate is, however, of limited value for my analysis. First, it does not yet sufficiently connect ‘manipulation’ in moral-political discussions with ‘manipulation’ in discussions on free will agency, even though both kinds of discussions have an interest in the same notion of ‘autonomy’.⁴⁷ It also seems to analyse manipulation by tacitly taking ‘autonomy’ as *the* model of agency without exploring the question of whether certain notions of agency imply certain notions of manipulation, more specifically: whether the distinction between negative and positive freedom has any consequence for analysing manipulation. My claim is that it does. I think that it is possible to distinguish between two types of manipulation. One of them is relevant to both negative and positive notions of freedom. It is threatening to autonomy, but not in any distinctive way (it is threatening to negative freedom as well). For now, I am interested in the second type of manipulation that I think can be specifically linked to positive notions of autonomy and for which the ‘influencing machine’ is the appropriate metaphor. Below I will first discuss Berlin’s criticism on positive freedom because it provides some clues to distinguishing what specific type of manipulation can be linked to positive freedom. Next, I will make a proposal for a *thick* notion of manipulation, which I will use for distinguishing *two types* of manipulation, one of which is distinctive for the debate on autonomy.

Berlin’s distinction between two concepts of freedom has been extensively discussed ever since the publication of ‘Two Concepts of Liberty’. But his criticism of positive freedom has been obfuscated somewhat because various commentators have focused on his distinction as a proposal to analyse freedom as such. Some have therefore responded by claiming there is only one concept of freedom, while others have claimed that even more concepts should be distinguished.⁴⁸ Arguments like these are part of an effort to elucidate the conceptual relations between negative and positive freedom and, in this way, gain more insight into the general structure of the notion of freedom. Here I will draw on Santoro’s suggestion that Berlin had a different purpose in distinguishing between both concepts of freedom. Berlin was not interested in the exact conceptual relations

45 See e.g. Thalberg (1978); Lehrer (2003).

46 Influential contributions to this debate are e.g. Rudinow (1978); Goodin (1980); Faden and Beauchamp (1986); Noggle (1996); Coons and Weber (2014a). In his overview of the debate on manipulation, Noggle (2018:4-5) mentions Faden and Beauchamp (1986) as ‘one of the earliest sustained philosophical discussions of manipulation’.

47 Noggle, for example, in his article ‘The Ethics of Manipulation’ for the online *Stanford Encyclopedia of Philosophy*, recognises that the term ‘manipulation’ is also central to discussions on free will (in the context of a determinist world). He refers to it as ‘global manipulation’ in contrast to ‘ordinary manipulation’, which is the subject of his own article. While he acknowledges that ‘few people have explored the connections’ between both forms of manipulation and that ‘it is still worth wondering about the relationship between them’ (2018:4), he does not discuss their relationship. A similar decision is taken in the volume Coons and Weber (2014a): see Coons and Weber (2014b, note 5).

48 See Santoro’s discussion on the reception of Berlin’s essay (2003:33–34).

between them. He even conceded that they ‘start at no great logical distance from each other’ (1958/ 2002:35–36). Instead, he was concerned with a crucial difference between both concepts that was to be found at the level of analysing not so much freedom but the underlying view on human beings.⁴⁹

To Berlin, it is important how notions of freedom are used as *political* notions. In his view, political freedom is already adequately defined as non-interference (‘negative freedom’). Political theory has no need for a notion of subjectivity that introduces extra requirements for what it means to be politically free.⁵⁰ The *political* notion of negative freedom may, of course, relate to a notion of subjectivity. But it does not imply any *additional* requirements for being free: to be free is simply not to be interfered with. This was, as we have seen, already Hobbes’s strategy. Hobbes did not accept any further subjective conditions for the will, as he made clear in a polemic against Bramhall: ‘I acknowledge that this liberty, that I can do if I will; but to say I can will if I will, I take to be an absurd speech’ (Chappell 1999:16). When Frankfurt says that freedom consists in being ‘free to want what he wants to want’ (1971:15), this would be for Hobbes an ‘absurd speech’ and, I assume, for Berlin as well.⁵¹ Opponents of positive freedom, however, as Berlin understands them, will claim that subjectivity implies extra conditions for determining what political freedom is about. The reason for this is that they have a radically different conception of subjectivity, one that presupposes ‘man divided against himself’ or the ‘splitting of personality into two: the transcendent, dominant controller, and the empirical bundle of desires and passions to be disciplined and brought to heel’ (1958/ 2002:181). In short: they are committed to *hierarchical* notions of subjectivity.

Berlin criticises this hierarchical notion of subjectivity because it allows adherents of positive freedom to slide easily into the justification of paternalistic practices. How pursuing one’s own autonomy transforms into paternalism, in his view, may be reconstructed as follows. Someone who pursues his Real Self (his ‘higher nature’) will want to establish first what this Real Self is. In this process, she will find certain reasons for pursuing what she discovers as her Real Self. She may conclude that *her* reasons for what is right and what is a Real Self also obtain for others: ‘For if I am rational, I cannot deny that what is right for me must, for the same reasons, be rights for others who are rational like me’ (Berlin 1958/ 2002:191). She will also observe then how others do not pursue these same reasons ‘because they are blind or ignorant or corrupt’ (1958/ 2002:179). A decisive step into the domain of paternalism is taken once this person

49 See Santoro (2003: 35–39).

50 Non-interference may also be interpreted in subjective terms because it still needs, as some have argued, the notion of a ‘want’ because how else will you define ‘non-interference’? Non-interference is when you can do what you want to do. But in a different context, Berlin makes clear that this is not what he has in mind: ‘the definition of negative liberty as the ability to do what one wishes – which is, in effect, the definition adopted by Mill – will not do’ (2002:186). Instead, we should understand ‘non-interference’ in a more general sense: an area or domain where people’s actions are not restricted.

51 Yet another phrase of Frankfurt that Hobbes would dislike: ‘to will what he wants to will, or to have the will he wants’ (1971:15).

understands her observation as justifying her role as ‘social reformer’: ‘if my plan is fully rational, it will allow for the full development of their ‘true’ natures, the realisation of their capacities for rational decisions, for ‘making the best of themselves’ – as a part of the realisation of my own ‘true’ self’ (1958/ 2002:193). The social reformer may then go on to ‘ignore the actual wishes of men or societies, to bully, oppress, torture them in the name, and on behalf, of their ‘real’ selves’ (1958/ 2002:180) because ‘I cannot ask their permission or consent, because they are in no condition to know what is best for them’ (2008:197). Even more so, these people would also pursue her reasons if they had been ‘more enlightened’. This thought makes it easy for her ‘to conceive of myself as coercing others for their own sake, in their, not my, interest’. This brief account takes the perspective of a single person, but we might as well imagine how a group of people adopts a similar reasoning and see themselves as social reformers, the liberators of others because what the Real Self means, does not need to be limited to some personal goal: ‘the real self may be conceived as something wider than the individual (as the term is normally understood), as a social ‘whole’ of which the individual is an element or aspect: a tribe, a race, a Church, a State, the great society of the living and the dead and the yet unborn. This entity is then identified as being the ‘true’ self which, by imposing its collective, or ‘organic’, single will upon its recalcitrant ‘members’, achieves its own, and therefore their, ‘higher’ freedom’ (2008:179).⁵² Once a group of people (in whatever sense: political, religious or otherwise) begins to pursue the ideal of the Real Self, this might be the start as well of an illiberal regime. However, the main point of Berlin’s criticism is that as Santoro (2003:36) summarises it: ‘pursuing positive freedom irredeemably compromises negative freedom’. Paternalistic policies will have the consequence of limiting people’s options, ‘for their own sake’, and in this way diminishing what Berlin considers as the right kind of liberal freedom: negative freedom.

Apart from his criticism of positive freedom as a justification of paternalism, Berlin’s analysis provides yet another insight into positive freedom, not at the level of how it justifies paternalistic practices, but at the level of what it assumes about human subjectivity, namely that people are vulnerable to being subjected by their paternalistic rulers to *unconscious* forms of paternalism.

What is typical about paternalism is not necessarily a certain type of action, but any action, whatever form it takes, that is employed to promote the wellbeing of certain people, but in a way that interferes with their options (negative freedom, Berlin would say) without their consent.⁵³ Now the paternalistic rulers may choose to impose their policies in different ways: they may simply take recourse to coercion (violence) or

52 Interestingly, the criticisms that communitarians, such as Sandel [1992] (1998) raised against the notion of autonomy as Rawls used it (in his Theory of Justice), might – according to Berlin’s analysis here – be interpreted as still following the pattern of autonomy!

53 For this definition of paternalism, I draw on the suggestions made by Dworkin (2017:1): ‘Paternalism is the interference of a state or an individual with another person, against their will, and defended or motivated by a claim that the person interfered with will be better off or protected from harm’.

they may, which Dworkin includes in his definition of paternalism, engage in rational persuasion and defend their policies with ‘a claim that the person interfered with will be better off or protected from harm’ (Dworkin 2017:1). Berlin repeatedly makes clear that the paternalistic rulers perceive the ‘unfree’ as ignorant (irrational, desire-ridden). But to what extent does Berlin agree with this perception? It is one thing to argue, based on negative freedom, against the justification of paternalistic policies, but it is another thing to argue that people have the abilities to recognise the gap between their negative freedom, both ideally and factually, and what their rules do to diminish this freedom. Berlin would agree, I assume, that people are able to recognise that the use of violence by paternalistic rulers is an infringement upon their negative freedom. But to what extent are they able to recognise that the reasons paternalistic rulers invoke for their paternalistic policies are consistent with positive freedom but not with their negative freedom? It would require that people have had the opportunities to develop their rational skills sufficiently.

Interestingly Berlin seems to recognise, besides rational persuasion and coercion through violence, yet another strategy of imposing paternalistic policies on ‘recalcitrant human beings’: manipulation. The term ‘manipulation’ is used by Berlin several times in his text, but – admittedly – not always in the same sense and not with the aim of systematically analysing manipulative actions. Nevertheless, he does seem to recognise manipulation as a distinct strategy. At some point (in section III, *The retreat to the inner citadel*), Berlin, rather than showing how paternalistic policies on the basis of positive freedom violate negative freedom, tries to lay bare the inner contradictions in the plea for such policies: ‘if the essence of men is that they are autonomous beings [...] then nothing is worse than to treat them as if they were not autonomous, but natural objects, played on by causal influences, creatures at the mercy of external stimuli, whose choices can be manipulated by their rules, whether by threats of force or offers of rewards’ (1969/ 2002:183). Later in his argument, it appears that Berlin has in mind certain techniques that turn people into ‘natural objects’ and ‘creatures at the mercy of external stimuli’. I quote two crucial passages:

‘All forms of tampering with human beings, getting at them, shaping them against their will to your own pattern, all thought-control and conditioning, is, therefore, a denial of that in men which makes them men and their values ultimate’ (1958/ 2002:184).

‘If the tyrant (or ‘hidden persuader’) manages to condition his subjects (or customers) into losing their original wishes and embracing (‘internalising’) the form of life he has invented for them. He will, no doubt, have made them *feel* free [...]’ (1958/ 2002:186).

Both passages reveal that Berlin was aware of techniques of influencing people that were different from coercion and rational persuasion. In addition to ‘thought-control’ and ‘conditioning’, he also mentions, in a note, ‘the techniques of hypnosis, ‘brainwashing’, subliminal suggestion and the like’ (1958/ 2002:184). (Berlin does not use one name to refer to these techniques as one category of manipulation, but we might call them, for now at least, techniques of ‘hidden persuasion’.) In the second quote, the phrases ‘hidden persuader’ and ‘customers’ indicate that Berlin, very probably, was familiar with the bestseller *The Hidden Persuaders*, published in the United States in 1957 (just one year before Berlin published his ‘Two Concepts of Liberty’), in which Vance Packard tried to show how advertisers tried to influence consumers into buying their products by employing techniques that, rather than addressing their *conscious* rational abilities, tap into their *unconscious* desires.⁵⁴ In this respect, ‘hidden persuasion’ differs from either violence or rational persuasion. People are not ignorant about the use of violence against them. They may be able to unmask the reasons in support of paternalistic policies. But people are at least ignorant, by definition, about the effects of hidden persuasion.

It looks as if Berlin intends to strengthen his core criticism against positive freedom – its function in justifying paternalistic practices – by invoking the threats of the techniques of hidden persuasion because rulers may use these techniques to even *conceal* their use of paternalistic practices. In this way, Berlin claims, the rulers create ‘the very antithesis of political freedom’ (1:186), which, in Berlin’s view, is negative freedom. But is that so obvious? In recognising this threat of manipulative techniques Berlin arrives at an awkward point in his analysis of positive freedom. He recognises the effects of manipulative techniques on people’s minds. It follows, for example, from his brief, passing comment on the psychology of Kant, the Stoics and Christians to the extent that it makes the assumption ‘that some element in man – the ‘inner fastness of his mind’ – could be made secure against conditioning’, because he remarks that conditioning by hypnosis, brainwashing and subliminal suggestion ‘has made this a priori assumption, at least as an empirical hypothesis, less plausible’ (1958/ 2002:184, note 2). And yet, if the notion of positive freedom is rejected, not just as a moral-political ideal, but also as a view of human subjectivity as having a hierarchical structure, it is not obvious how else to conceptualise the effects of hidden persuasion on people’s inner world. Coole denies that negative freedom, as Berlin favours it, can conceptualise the ‘new techniques of power’ and therefore concludes that ‘the criteria of negative liberty and the forces it confronts [...] would have to be rethought’ (2013:211).⁵⁵ After all, the view of human

54 Coole (2003) makes this suggestion.

55 As part of her criticism of Berlin’s analysis, Coole claims that there is ‘a good deal of difference between theories that would liberate a self held in thrall to its passions on behalf of deluded ideals of moral perfection or authentic subjectivity and one that recognises that empirical individuals might be psychologically manipulated as a deliberate ruse of power in pursuit of certain interests’. I am not sure what theory she hints at that ‘recognises that empirical individuals might be psychologically manipulated’. In her article, she mentions Adorno and Horkheimer, also Foucault. But the force of her criticism, Berlin would respond, depends on the plausibility of hierarchical notions of subjectivity.

subjectivity implied in negative freedom makes no distinction between a will that originates in a Real Self and a will that originates in an inauthentic or, for that matter, *manipulated* Self.

The inconsistency in Berlin's analysis of positive freedom may nevertheless bring us closer to understanding the special link between positive freedom and certain forms of manipulations. Below I will offer an analysis of manipulation that will shed light on this link. It includes a proposal to define manipulation and to distinguish between two types of manipulation, one of which is compatible with both notions of freedom, while the other can be expressed only within the conceptual framework of positive freedom.

Theorists on manipulation generally agree that manipulation is some form of influence that is different from either coercion or persuasion.⁵⁶ Roughly two distinct (families of) views can be distinguished on what demarcates manipulative from non-manipulative influencing.⁵⁷ According to one view, what is essential about manipulation is *deception*.⁵⁸ According to another view, manipulation is 'a process of influence that necessarily either bypasses or subverts the rational capacities of the person whose behavior is being influenced' (Gorin 2014:52).⁵⁹ Theorists sometimes argue against one of both views.⁶⁰ But I think what both views have in common is that they understand manipulation as interfering with people's *rationality*. The two views may then be understood as suggesting two different types of actions to undermine rationality: *reason-deceiving* actions and *reason-bypassing* actions. I will therefore refer to both views by calling the first the *reason-deceiving* view and the second the *reason-bypassing* view. Taking their common idea as a starting point, I propose a preliminary, broad definition of manipulation:

Manipulation is an influencer's knowing use of techniques that activate one or more of the influencee's epistemic dispositions and which result in the influencee adopting false beliefs, which in their turn induce the influencee to perform certain actions while the influencee is unconscious of having false beliefs.⁶¹

56 See, e.g. Noggle (2018:2): 'Manipulation is often characterized as a form of influence that is neither coercion nor rational persuasion'.

57 See Coons and Weber (2014b:10), who distinguish 'two predominant ways of characterizing manipulation: as a kind of *deceptive* non-coercive influence and as a kind of *nonrational* influence'. Also, see Noggle (2018), who makes a similar distinction.

58 For references: see Coons and Weber (2014b:10, note 6). Noggle (2018) discusses this branch of views by calling them 'trickery'.

59 This view is discussed, for example, by Noggle (2018). For references to scholars who subscribe to this view: see Gorin (2014:59–60, note 1). Also see Mele (1995); Todd (2013); Sunstein (2016).

60 Arguing against deception views is, for example, Cohen (2017). Gorin, on the other hand, argues against the Bypass or Subvert View. But, on closer view, he does not seem to deny that reason-bypassing manipulations may exist but instead claim that (what I call) reason-deceiving are more common.

61 Perhaps I should add: '*non-coercive* use'. The proper way to distinguish manipulation and coercion, I think, is to understand coercion as 'coercion uses threats, which involve changing the costs of selecting certain options' (Wilkinson 2012:5). For my analysis of manipulation, I am mainly interested in the notion of rationality and, therefore, in tracing the differences between persuasion and manipulation in relation to people's rational capacities. I think the boundary between coercion and manipulation has less to do with the notion of rationality and more with the notion of power. Therefore, I choose not to explore the exact differences between manipulation and coercion.

Although I am sure that my definition has its faults, it will allow me to include both reason-deceiving and reason-bypassing actions. I will start with three comments on my definition and then move on to providing narrow definitions of manipulation.

First, what may be striking about this definition is that it does not explicitly mention ‘rationality’. The reason for this is that any definition that starts from the idea that manipulation undermines people’s rationality still should specify what exactly this means. For my broad definition, I have chosen to understand what is distinctive about manipulative actions in terms of their *impact* on people’s rationality, which is to interfere with it, and to specify this impact in terms of the false beliefs which people adopt as a result of these actions. I have also assumed that the impact of manipulative actions is mediated through their effect on people’s ‘epistemic dispositions’. What these dispositions mean will be explained when I turn to the narrower definitions of manipulation.

Second, I have chosen to use the phrase ‘techniques’, but this is not essential to my definition. What suffices for any action to count as manipulative, according to this broad definition, is that it succeeds in getting people to adopt certain false beliefs. But I do believe that manipulative actions will usually involve some ‘technique’, in the sense that influencers must somehow have learned to master these actions. While they do not need to have the technical knowledge of how exactly the technique works (they might understand it only in an abstract way), at least they need to know what impact the technique will have on the influencee, and what to do themselves for this technique to have that impact, i.e. how to ‘start’ or ‘activate’ that technique.

Third, my definition is intended to obviate, for the moment at least, the issue of the permissibility of manipulation. It is supposed to be neutral, for example, to whether the manipulator has good or bad intentions.⁶² Although manipulators are usually thought to strive for their own interests, I think my definition allows understanding manipulation as a form of paternalism (‘for their own sake’) and even as what we would consider ordinary or benign forms of ‘manipulation’ as in letting children believe in, for example, Santa Claus or the Tooth Fairy. If getting people to adopt false beliefs is *in principle* unjustified, then, of course, this definition fails to be neutral.

Now that the broad definition has been clarified to some extent, the next step is to specify it into two distinct, narrow definitions, one which covers reason-deceiving manipulation, another which captures reason-bypassing manipulation. The broad definition simply assumes some ‘causal’ connection between manipulative actions and their impact. In shifting from the broad to the two narrow definitions, I will try to provide a more specific characterisation of manipulative actions. It will help to explain

62 If manipulation as such is unethical or immoral, this is not, in my view, because the manipulator’s intentions are immoral. Manipulation can be used for good intentions, as for paternalistic intentions (unless one has reasons for judging that paternalistic intentions are immoral) in the same way that logical reasoning can be used for bad intentions.

the distinction between the two types of manipulation. I will start with a narrow definition of reason-deceiving manipulation:

Reason-deceiving manipulation is an influencer's *conscious* use of techniques that activate one or more of the influencee's epistemic dispositions and which result in the influencee adopting false beliefs, which in their turn induce the influencee to perform certain actions, while the influencee is unconscious of having false beliefs *as false*.

The paradigmatic case for this type of manipulation is *lying*. The influencer will present certain false facts as true, knowing that it is likely to motivate the influencee to perform some action. This may be facts about the options that are available to the influencee, or facts about what will happen when the influencee chooses certain options, or perhaps, often in cases when the influencer and influencee know each other well, certain facts about the influencer herself, about her motives or feelings. Lying is a well-known topic for philosophical analysis. But lying presents only one kind of deception, one that is specifically linguistic. Pretence offers another, less well-known case of deception, a non-verbal one.⁶³ Both verbal and non-verbal lying, as do truthful speaking and doing, appeal to people's 'epistemic dispositions', which, in the case of reason-deceiving manipulation, refer to their capacities to deliberate based on facts they know. But while truthful speaking and doing involve true facts, lying introduces false facts. In this sense, lying interferes with people's rationality.

For the second type of manipulation, I suggest the following narrow definition:

Reason-bypassing manipulation is an influencer's knowing use of techniques that activate one or more of the influencee's epistemic dispositions and which result in the influencee adopting false beliefs, which in their turn induce the influencee to perform certain actions, *while the activation as such of these dispositions is something the influencee is unconscious of, and while the influencee is unconscious of the false beliefs*.

63 While the phenomena of non-verbal deception – such as pretence (*simulatio*) and the deceiver (the *fallax* as opposed to the *mendax*, the liar) – have generally been ignored by philosophers, interestingly in the sixteenth and seventeenth centuries, many theologians, philosophers and others have been interested in them. See, e.g. Van Houdt (2002), Snyder (2009).

What I have in mind here is what Faden and Beauchamp have called ‘psychological manipulation’: ‘a person is influenced by causing changes in mental processes other than those involved in understanding’ (1986:355).⁶⁴ While both reason-deceiving and reason-bypassing actions can be said to have an epistemic impact, they are different, therefore, in *what kind* of ‘epistemic dispositions’ they activate: as reason-deceiving actions activate people’s normal capacities for deliberative reasoning, reason-bypassing actions activate certain *non-deliberative* epistemic dispositions. These dispositions are somehow cultural, psychological or even psychophysical in nature. What is distinctive about them is that the influencees have no direct (i.e. unreflected) knowledge of them. But in calling these dispositions ‘epistemic’ I wish to emphasise their epistemic dimension, i.e. the extent to which these dispositions can elicit an impact on the influencee that can be made sense of in epistemic terms, i.e. as ‘beliefs’. Whether influencees can have knowledge, reflective or otherwise, is what I will explore in the next chapters.

The category of reason-bypassing actions covers two different subtypes of manipulation: the paradigmatic case for the first is *seduction* – an example of which might be wearing cologne on a date⁶⁵ – while for the second, it is *brainwashing*, to which we might add: hypnosis and subliminal influencing. The differences between both subtypes might be explained in different ways. The definition of seduction might, for example, be supplemented with: ‘... *unconscious of the false beliefs as false*’. It would suggest some similarity between cases of pretence and cases of seduction. And I do think it is difficult to distinguish such cases. The definition of brainwashing and other cases might be supplemented with: ‘... *unconscious of the false beliefs, not just as false, but of the beliefs as such*’. It would suggest the possibility of unconscious beliefs that nevertheless have the potential to induce people to perform certain actions. Yet a different way of drawing a boundary between both subtypes is claiming that influencees might still resist seductive actions – wearing cologne is usually supposed to have a relatively mild effect (the fragrance may let us *feel* seduced but does not force us to *be* seduced) – while they have no option to escape brainwashing or hypnosis. If any form of manipulation has

64 I think that this is also what Raz (1986:377–378) had in mind with his definition: ‘Manipulation, unlike coercion, does not interfere with a person’s options. Instead, it perverts the way that person reaches decisions, forms preferences or adopts goals’. Barnhill (2014:53–54) calls it ‘direct manipulation of a person’ as opposed to ‘manipulation of a situation’. In making this distinction, she also refers to Faden and Beauchamp. However, her explanation of what it means to directly manipulate a person is, I think, less clear than theirs. Faden and Beauchamp also mention two categories of actions that consist in tampering with the circumstances that people are situated in: (a) *manipulation of options* ‘in which options in the person’s environment are modified by increasing or decreasing available options or by offering rewards or threatening punishments’ and (b) *manipulation of information* ‘in which the person’s perception of these options is modified by nonpersuasively affecting the person’s understanding of the situation’ (Faden and Beauchamp 1986:355). But my suggestion would be to categorise these forms as belonging to the type of manipulation that appeals to people’s rational abilities.

65 The example of wearing cologne is suggested by Todd (2013), but he dismisses it as a case of manipulation.

captured the imagination of philosophers and provoked their worries about autonomy or moral responsibility, if any form of manipulation has fuelled philosophical anxieties about people being ‘puppets on strings’ or, to keep with the metaphors of technological progress, ‘automatons’, this has been the second subtype of manipulation: acts of manipulation that bypass or subvert people’s rationality (whatever that exactly means because we still need to find that out), especially those that are forceful.⁶⁶

The distinction between the two main views on manipulation, as I mentioned earlier – deception (‘trickery’) versus bypassing reason – is to some extent accepted in the debate on manipulation. I suggested that ‘interfering with rationality’ is distinctive for manipulative influencing and that we should categorise manipulative influencing in terms of *how* they interfere with people’s rationality (either in a deliberative or non-deliberative way). But this does not necessarily represent a standard view in the debate. Therefore, the two types of manipulation for which I proposed a definition do not neatly correspond with the two main views in the standard debate. One reason for this may be that each view is analysed on its own, without exploring its links with the other view. Or the reason might be that rationality, or some aspect of it, is frequently included in discussions on manipulation, but usually without exploring systematically or in depth what the relation is between rationality and manipulation.⁶⁷ By taking rationality as a starting point, I arrived at a different distinction: What I have discussed as the first type of manipulation represents an old phenomenon: lying. In my view, what many theorists discuss under the heading of ‘deception’ are cases of the second type of manipulation: cases therefore that involve examples of the activation of what I would call a *non-deliberative epistemic pattern*. Interestingly the debate on manipulation does not present itself as a continuation of older debates on lying but has started all over, taking ‘manipulation’ as its core notion. What may explain this, I think, is the discovery, due of course to ongoing insights in psychology, that people can be influenced in all kinds of ways that go unnoticed by them. This is sometimes understood in the debate on manipulation as implying that manipulation is covert (which is denied by still other theorists). But this, as I will try to explore in the next chapters, depends on how rationality is understood.

66 See, e.g. the examples mentioned by Frankfurt – a certain Black who manipulates a certain Jones: various tools are suggested: a potion, hypnosis and some sort of tool that intervenes directly in Jones’s brain and nervous system (1969) – and by Mele – a certain Beth who is manipulated by brainwashing (1995). In the next chapter, I will return to such examples.

67 For an exception: Fischer (2017). Other discussions are usually uninformative. For example, Coons and Weber (2014b) distinguish a type of ‘*nonrational* influence’. But from this discussion, it becomes not clear what warrants this qualification. Gorin explicitly discusses cases of manipulation in terms of whether they bypass rationality. But I think his discussion, in the end, is unsatisfactory too. Gorin claims that most cases of manipulation involve an appeal to people’s rationality and therefore cannot be described as bypassing or subverting people’s rationality. But he does not recognise that even in these cases, rationality is interfered with, namely that people end up with false beliefs. And he also does not recognise that even cases that ‘entirely bypass’ people’s rationality still assumes that people have rational capacities, without which brainwashing would not even be possible.

The two types of manipulation pose a threat to people's freedom in both senses: negative and positive freedom. Adherents of both notions can explicate the specific threat that the first type, reason-deceiving acts of manipulation, poses to people's freedom. For this, a notion of deliberative rationality suffices, while their theory of justice merely needs to assume that people perform truthful (speech)acts.⁶⁸ This is different from the second type: reason-bypassing acts of manipulation. For adherents of negative freedom, it suffices for an act to count as free, and therefore also as non-manipulated, if it is not interfered with (in Berlin's sense). They do not require any additional conditions for a person's will that underlies her action. Consequently, however, they do not have the theoretical resources to explicate the dangers that acts of the second type pose to people's freedom. Berlin, as we have seen, even denies that whatever problems issues like hypnosis raise do not pose a problem *for freedom*. But Berlin too can be recognised to be, in the end, ambiguous about analysing these problems.⁶⁹ Only notions of positive freedom provide the theoretical resources to acknowledge the threat that manipulative actions of the second type pose to people's freedom. They assume that people's subjectivity is hierarchical. It allows for the possibility that unconscious forces are at work. Even if notions of positive freedom may explain how, however roughly, people's inner freedom may be undermined, this does not mean that these notions also provide the conceptual recourses for a solution. This will be examined in the next section.

2.4. Reflective epistemic autonomy

In the previous section, I made a distinction between two types of manipulation. Only actions of the second type – actions that bypass or subvert people's rationality – count as *non-deliberative* influencing actions, which is the subject of my research question, while actions of the first type (untruthful actions) do not. I will therefore pursue my discussion by focussing on the second type. This section starts with a proposal for understanding how manipulation bypasses (or subverts) people's rational capacities and then presents what I call *reflective epistemic autonomy*. It provides a possible answer to how the liberal notion of responsibility may deal with manipulation.

The second type, those acts of manipulation that bypass (or subvert) people's rational capacities, are a source of worries for adherents of positive freedom because it seems to undermine their autonomy. Is it possible for people to maintain their

68 For the early contractarian theorists, the notion of (negative) freedom turned on trust. Any untruthfulness might endanger the social contract and, therefore, each person's freedom. For Grotius: see Bouchilloux (1997). For Hobbes and Locke: see Schröder (2018). Also, see Santoro (2003), who links the issue of trust to the problem of *akrasia*.

69 Although the possibility of a hidden persuasion was an additional reason for Berlin to reject definitions of freedom in terms of subjectivity, his position about its implications for understanding freedom may have been more ambiguous. Of five examples he provides of what is both negative and positive freedom, two of them are 'surely internal', as Bellamy observes: 'mass hypnosis of custom or organised propaganda' (Bellamy 2000:23).

autonomy? And if so, how? But this type raises the question of what it means that rational capacities are bypassed (or subverted). Unfortunately, as I have suggested already in the previous section, theorists on autonomy provide little explanation. They rely on their readers to understand what it means that cases of brainwashing or hypnosis are examples of manipulation and, by consequence, of bypassing (or subverting) people's rational capacities.⁷⁰ What manipulation is *not* doing is affecting the rational capacities, at least not in the sense that manipulation dissolves (in whatever sense) or temporarily interrupts rational capacities. What manipulation *is* doing, in my view, is that it hides from sight, metaphorically speaking, what people – under non-manipulative conditions – would be able to judge critically, using their rational capacities. My proposal is, therefore, to put it yet differently, that manipulation, instead of appealing to 'mechanisms' of reasoning, is *activating behavioural mechanisms* in ways that bypass people's self-knowledge and *therefore* bypasses people's capacity to critically judge the activation of that behavioural mechanism (whatever we take this to be: some psychological or perhaps some neurological mechanism). If this proposal is acceptable, then it shifts the discussion of manipulation away from a debate on how people choose their 'will' or their 'first-order desire' towards a debate on how people determine their beliefs. What would be at stake, after all, would be the question of whether people have control over their beliefs. If manipulation bypasses our behavioural mechanisms by bypassing our cognitive systems (or belief systems), this would undermine our autonomy. If, on the other hand, it could be made plausible that people, despite the influence of manipulation, can still keep control over their beliefs and therefore over their behavioural systems, perhaps this shows how manipulation is less threatening to autonomy than is usually assumed.

In the remainder of the section, I will discuss reflective epistemic autonomy. As illustrative for this notion, sometimes this passage (in part or even more extensively) is quoted:⁷¹

'our capacity to turn our attention on to our own mental activities is also a capacity to distance ourselves from them, and to call them into question. I perceive, and I find myself with a powerful impulse to believe. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn't dominate me and now I have a problem. Shall I believe? Is this perception really a *reason* to believe? [...] The reflective mind cannot settle for perception and desire, not just as such. It needs a *reason*. Otherwise, at least as long as it reflects, it cannot commit itself or go forward.' (Korsgaard 1996a:93)

70 Raz, for example, does not explain his claim that manipulation 'perverts the way [a] person reaches decisions, forms preferences or adopts goals' (1986:377–378).

71 See e.g. Owens (2000:9-11); Moran (2001:142); Engel (2002:14-15); Shah (2003:478-479); Kornblith (2012:109-110); Doris (2015:18).

According to this argument, even if people are not able to decide what they believe at will because they have these beliefs already as a result of natural or social processes of belief formation, at least they can reflect on ('to distance ourselves from') the beliefs they find themselves having.⁷² The argument is not unique for Korsgaard, the author of this passage. It belongs to a family of views that attribute an important role to reflection.⁷³ In discussing reflective epistemic autonomy, I will refer to Korsgaard as its representative, not only because her argument serves as a point of reference in the doxastic debate, but also because her argument is part of a more general view she has on freedom and which connects the determinist and doxastic debates.

It is not easy to spell out Korsgaard's argument. She mentions the possibility of reflectively gaining control over beliefs, as the above quote shows, but in further discussions, she focuses on explaining gaining reflective control over the will. Her explanation of control over belief remains concise. How it works is only hinted at.⁷⁴ Instead, I will turn to Richard Moran (2001), who is concerned with understanding introspection primarily as self-knowledge, which we should understand here not as *personal* self-knowledge, but in an epistemological sense: as the knowledge that has a distinctively *first-person perspective*.⁷⁵ Moran's ambition is to explain how belief formation, understood as an indirect process, is a matter of decision, but a decision that is up to me: it is *my* decision. Therefore, he is concerned with showing the distinctive importance of the first-person perspective as contrasted to the third-person perspective on assessing beliefs. Even if Moran is not concerned with freedom primarily, he does link his notion of self-knowledge with a notion of free belief formation because understanding freedom requires understanding self-knowledge. He claims that having knowledge with the right sort of first-person perspective constitutes the kind of epistemic agency that is required for freedom.⁷⁶ In his discussion, he also provides an interpretation of Korsgaard's view. Later in this section, I will suggest how reflective will, which is Korsgaard's focus, and reflective belief, in Moran's view, may be viewed as connected. I then will also compare the Korsgaard-Moran reflective model of introspection with Frankfurt's reflective model.

72 The argument may be presented in various ways. Garvey, for example, chooses to present it as the 'Character-Choice Theory' (2008:162-163) while Engel discusses it under the heading of 'freedom of belief and the power of reflection' (2002:13-22). The central point remains the same: we do not have the power, prospectively as it were, to be the origin of our beliefs, but we do have the power to control our beliefs retrospectively.

73 The reflective view is also attributed to Kant, Hampshire, Moran and Sosa (see Kornblith 2010), to Velleman (see Wallace 2001), to Hare, Smith, McDowell, Gibbard, Wedgwood (see McHugh 2013), to McDowell, Burge, Smith, Brandom (see Engel 2002), to Hampshire (Moran 2001).

74 To my knowledge, Korsgaard has not yet fully participated in the doxastic debate.

75 Also, see Baker (1998), who sets up her argument for the first-person perspective more explicitly against naturalistic views.

76 In my presentation of Moran's argument, I will therefore take some liberty in suggesting 'freedom' in places where Moran is concerned with the right kind of first-person perspective.

The reflective formation of belief in Moran's view will be discussed as proceeding in four subsequent stages even if Moran himself does not make this distinction in exactly this way:

1. Stepping back from certain beliefs.
2. Double suspension of these beliefs.
3. (Deliberative) Reasoning about these (and other) beliefs.
4. The avowal of certain beliefs.

The first two stages will be discussed briefly. The last two stages require some more explanation.

We will start with a simple example. Someone is looking out of the window, distractedly, and sees it is raining. Suddenly, she becomes aware of what she is looking at and thinks, or perhaps mumbles: 'It is raining', as an expression of her belief that it is raining. The mental process that was going on in her mind can be constructed in terms of a transitive consciousness or, to use different phrases, in terms of first- and second-order beliefs. We might say that, in the first moment, she already had the belief that it was raining but was unconscious about it: a first-order belief. Then she reflected on what she was looking at and acquired a belief about the first-order belief, one that makes this belief conscious. Moran believes that the reflective step from first- to second-order belief does not yet capture the first-person perspective of reflection.⁷⁷ What does constitute a step in that direction is when people realise what commitments are involved in having beliefs. Thinking (or saying) 'it is raining' would express that a person believes it is raining. Therefore, if she would say, 'I believe it is raining', it seems nothing would be added. Moran disagrees. He claims that in such cases, two different types of commitments are involved:

'On the one hand [...] it is not an open question for me whether it is raining or not. At the same time [...] I must acknowledge myself as a finite empirical being, one fallible person in the world among others, and hence acknowledge that my believing something is hardly equivalent to its being true. And even when a person's fallibility is not the issue, anyone must recognize that his believing P is nonetheless an additional fact, distinct from the fact of P itself' (Moran 2001:74)

What Moran suggests here is a distinction between 'the fact believed and the fact of one's belief' (2001:62). It is a step into the direction of epistemic agency once people realise this distinction because it allows them to take a stance towards (a set of) their beliefs and, for example, view them as psychological states about themselves. In this context,

77 This claim is part of his criticism of Mellor (1977-78).

Moran employs the metaphor of ‘stepping back’.⁷⁸ Stepping back is what happens when people come to see the facts they believe as the facts *of* their beliefs: it turns their beliefs into a ‘psychic given’ (2001:14).

Stepping back still does not, by itself, establish the first-person perspective or, for that matter, freedom. But we will get there soon. The moment of stepping back from (of taking a distance towards) a belief opens the possibility of another stage in the first-person reflection: the suspension of that belief. It can be suspended in two different ways. One way involves calling that belief into question: the belief that before was taken to be true is now bracketed, i.e. its truth is suspended. Another way implies the suspension of its motivational force or, as Korsgaard describes it (also quoted by Moran): ‘now the impulse doesn’t dominate me’.

Once people have entered the stage of suspending some belief they have, ‘reflection demands a reason’ (2001:148). The suspension of belief forces a responsibility on people to respond to that belief in terms of reasons.⁷⁹ It raises the question: ‘What am I to believe?’⁸⁰ Whatever truth they assumed their ‘psychic given’ to have, is no longer evident. They now need to assess the reasons in favour of (or against) taking that belief to be true. There is no escape from this responsibility because ‘any response to that “given” can now be understood in terms of the person’s responsibilities, and hence as implying either “endorsement, permission, or disapproval,” or simply passive allowance’ (2001:148).⁸¹

Not any reasoning will be constitutive of people’s freedom. Moran is careful to distinguish between two different stances that people may take towards their suspended beliefs: either a theoretical or a deliberative stance.⁸² For Moran, the deliberative stance, the stance of agency, has primacy over the theoretical stance because this is what, in the end, really constitutes freedom from within a first-person perspective. The difference between both stances can be explained by looking at how people may answer two questions:

1. Is p true?
2. Does person X believe that p ?

78 This is too what Korsgaard hints at when she describes ‘our capacity to turn our attention on to our own mental activities’ and ‘to distance ourselves from them’ (1996a:93).

79 This is an epistemic responsibility in distinction to a volitional responsibility (Moran also uses the phrase ‘internal responsibility’ (2002:200). Also, see Engel for similar remarks on responsibility (2002:13). Distinguishing this stage of reflection again agrees with Korsgaard’s view: ‘The reflective structure of human consciousness sets us a problem. Reflective distance from our impulses makes it both possible and necessary to decide which ones we will act on: it forces us to act for reasons’ (1996:113). As I remarked before, Korsgaard chooses to focus on what this reflective distance implies for what people decide on action. But this passage also applies to beliefs in the sense that reflective distance also ‘forces us to believe for reasons’.

80 Korsgaard (1996a:93): ‘Shall I believe?’

81 Moran (2001:143): ‘the situation of reflective consciousness of some psychic given makes inescapable for the person a situation of decision, inescapable in the sense that anything the person now does will represent a choice of his, even if only by default.’

82 Moran also calls them ‘two kinds of responsibility’ (2002:198-203).

Suppose that p stands for some event in the world, for example: ‘a truck is coming down the street’. In such cases, the first question is world-directed because in answering this question, a person needs to attend to outward phenomena. Suppose a person is asked whether someone else believes that p , her answer does not necessarily entail an answer to the question of whether p is true. If she is asked, however, whether she herself believes that p , a question that is now self-directed, her answer to this question does entail the answer to the question whether p is true: ‘from within the first-person perspective, I treat the question of my belief about P as equivalent to the question of the truth of P ’ (2001:62:63). This is the Transparency Condition: the question of a person’s belief is transparent to the question of truth.

The Transparency Condition also holds in cases that p stands for some event in a person’s inner world, for example, a feeling of anger or betrayal. By taking a deliberative stance towards the question: ‘Do I believe that I feel anger?’ people shift to the question: ‘What am I to believe about my feeling anger?’ This is a deliberative question that is answered by determining what is true. If a person, however, takes a theoretical stance to the question: ‘Do I believe that I feel anger?’, she treats p (X feeling anger) as a psychological fact about a person who happens to be her, a fact also she is unaware of. The question would be settled, not by deliberation but by evidence. In such cases, the theoretical stance is really a third-person perspective because other people may take the same theoretical stance towards her, and they would have access to the same evidence. Both stances produce forms of reasoning that settle the truth of p , but these forms perform different functions. Theoretical reasoning explains, in a psychological sense, why some person believes she feels anger. The explanation may predict how this person will behave in certain decisions. On the other hand, deliberative reasoning is concerned with providing a justification for a certain belief and ends in the avowal of the conclusion that follows from deliberative reasoning.

Even if a person assumes responsibility in a deliberative way, this does not yet realise epistemic agency.⁸³ What concludes the reflective process towards adopting a genuine first-person perspective and thereby constitutes free belief is the avowal of belief.⁸⁴ To explain this, I will discuss the case of an akratic person that chooses to go into therapy.⁸⁵ This will allow me to rehearse once again the various notions of Moran’s argument. It will also give me the opportunity to interpret his argument as suggesting a path towards a practice of ‘holding responsible’ that requires neither coercion nor manipulation.

83 Also, see Reginster (2004).

84 This should explain motivation: see McHugh (2013), who calls this type of views ‘motivational internalism’.

85 This case brings together two themes that Moran uses himself at several points in his argument: the akratic person and the therapeutic context. Moran occasionally refers to the therapeutic context. On this, he comments: As Moran observes: (2004:12): ‘The connections between the philosophical and the therapeutic issues concerning self-knowledge are more important to the book’s aspirations than few explicit references may suggest.’

The reflective processes of the akratic person require a somewhat different analysis than those of the non-akratic person, as we will see.

The akratic person Moran himself discusses is the case of an addicted gambler.⁸⁶ We may also consider Frankfurt's unwilling drug addict as a case of an akratic person. Both addicts can enter, in some way, the first stage of reflection. One part of their history of addiction is that the addicts have consistently experienced addictive desire (for gambling, taking drugs). Another part is that they have been discontent about it and therefore tried, half-heartedly and unsuccessfully, to stop the desire. They know, therefore, that their history of addiction is also one of backsliding into an obstinate habit of addiction. We may imagine how both addicts follow through the first two stages of reflection, standing back and suspending, regarding their own history of addiction and particularly about their desire to take drugs. The first stage allows them to have the (second-order) belief that they have an addictive desire and another (second-order) belief that they are unhappy about the behavioural pattern of their addiction. In the next stage, they will notice that they are able to bracket the truth of these beliefs (one sense of 'suspension') but are unable to suspend the motivational force of the addictive desire (the other sense of 'suspension'). This is clearly a difference with the reflective processes of normal, non-akratic people.

The third stage of reflection requires that people take responsibility for their beliefs. For the addicts, it would mean that they should engage in reasoning on the origins of their addictive desire. In the standard case of reflection, people have certain beliefs or feelings which they can access reflectively. But the problem for akratic people is that they lack introspective access to a relevant part of their mental states (beliefs, feelings) that might give them a clue to some of their behavioural dispositions. As these are closed off from their introspection, it seems they are incapable of reasoning about their addictive desire and only have the option of being 'psychologically realistic' about themselves: treating their desire for gambling or taking drugs and their history of backsliding simply as so many merely psychological facts about themselves. Moran calls this state of mind one of resignation or acquiescence. But of course, this attitude is itself the mark of taking responsibility for one's beliefs, in this case adopting a theoretical stance towards oneself, 'for even resignation and acquiescence involve acceptance of a predictive reason about what I will in fact do' (2001:148). And so akratic people, such as Moran's gambler and Frankfurt's unwilling drug addict, too cannot escape epistemic responsibility.

The addicts are capable at least of taking a theoretical stance towards their addiction. But their reflective situation is nonetheless a hopeless one: they seem to have no other option than acquiescence. Deliberative reasoning is not available to them. There is yet another option. As long as their feeling of discontent with addictive behaviour persists alongside the stronger desire for addiction, the addicts may accept the feeling of

86 This is actually a case he borrows from Sartre. Moran does not treat the case of the gambler as explicitly from a therapeutic perspective as I do here.

discontent as their own, in the sense of Frankfurt's identification and Moran's avowal, and disapprove of the addictive desire. As this will not suspend the motivational force of the addictive desire, it brings them into a situation that both Frankfurt and Moran recognise as one of estrangement.

It does not even stop here. The addicts may decide not to resign in their situation of estrangement but go into therapy. The analyst may attribute to the analysand, the addict, some 'unconscious attitude of resentment' (2001:31), or perhaps a feeling of betrayal, envy, fear or anger that will explain some of his behavioural dispositions (his habits of addiction). But the attribution of such beliefs raises a problem for the addict:⁸⁷

'The person who feels anger at the dead parent for having abandoned her, or who feels betrayed or deprived of something by another child, may only know of this attitude through the eliciting and interpreting of evidence of various kinds. She might become thoroughly convinced, both from the constructions of the analyst, as well as from her own appreciation of the evidence, that this attitude must indeed be attributed to her. And yet, at the same time, when she reflects on [...] whether she has indeed been betrayed by this person, she may find that the answer is no or can't be settled one way or the other' (2001:85).

The unwilling addict may step back from his desire to gamble or take drugs, but he cannot step back from certain feelings, unknown to him, that supposedly explain his addiction. He will therefore experience 'a split between an attitude I have reason to attribute to myself, and what attitude my reflection on my situation brings me to endorse or identify with' (2001:67). Even so, now that he has learned in therapy about some unconscious belief (something like: 'I believe I feel anger at my parents for having abandoned me'), even if it is an external, non-introspective belief, it offers him the opportunity of new stages of reflection. The first two stages of reflection are fully available only to the non-akratic person. But the last two stages of reflection – reasoning and avowal – are available to both the akratic and non-akratic person even though they may acquire self-knowledge in different ways – the akratic person through therapeutic self-attribution, the non-akratic person through introspection. The attributed attitude allows him to take responsibility for it: to either accept the attributed attitude and attribute it to himself or to reject it. If the analysand accepts the attributed belief as one that is attributed correctly to him, this either expresses a theoretical or a deliberative stance, as it does it for the non-akratic person.

Although the accounts of Frankfurt and Moran show interesting similarities, they also differ. In Frankfurt's analysis, the unwilling addict may have no hope for freedom.

87 The feelings mentioned here are all examples used by Moran himself.

In Moran's analysis, he has. Explaining this also shows, I think, how taking a deliberative stance is not yet enough to achieve freedom but also requires the *avowal* of reasons. To engage in finding reasons is one thing, but avowing whatever conclusion follows from them is a very different thing. Avowal is not simply a theoretical exercise: simply recognising that certain reasons force you to accept a certain conclusion:

‘One must see one's deliberation as the *expression and development* of one's belief and will, not as an activity one pursues in the *hope* that it will have some influence on one's eventual belief and will’ (Moran 2001:94).

As I understand Moran, avowing a belief is not an isolated mental event, but itself a process that a person undertakes to embed her deliberation into the wider structure of her mental states: ‘a conscious belief enters into different relations with the rest of one's mental economy and thereby alters its character’ (2001:30-31). The difficulties in avowing a new belief are more obvious in the case of akratic people than in those of normal, non-akratic people. Whenever the akratic person avows what the analyst attributes to him, his avowal is always in danger of relapsing into a theoretical rather than a deliberative acceptance because there is a ‘crucial therapeutic difference between the merely “intellectual” acceptance of an interpretation, which will itself normally be seen as a form of resistance, and the process of working-through that leads to a fully internalized acknowledgment of some attitude which makes a felt difference to the rest of the analysand's mental life’. And Moran continues: ‘This goal of treatment, however, requires that the attitude in question be knowable by the person, not through a process of theoretical self-interpretation but by avowal of how one thinks and feels’ (2001:89-90). As the addict seriously engages in the process of ‘working-through’, the unwilling addict too may have hope of achieving freedom.

One merit of Moran's discussion is that it provides a hint of how we might think of social change in a morally relevant sense, even if Moran himself does draw attention to this. My discussion of the unwilling addict that chooses to go into therapy shows how the therapeutic context may provide a paradigm of social change. We need not think of therapy in terms of psychoanalysis, as Moran probably does. What is interesting about therapy as a paradigm for social change is that therapy apparently fulfils certain conditions, whatever they are, that allow the analyst to attribute to the analysand an unconscious attitude and allow the analysand to accept the attributed attitude. The phenomenon of implicit bias, of course, provides us with the case of an akratic person: the unknowing racist. So, we might think of the social psychologist as an analyst rather than a manipulator when she tells the unknowing racist that he suffers from an implicit bias.

2.5. The mentalist roots of reflective epistemic autonomy

In the previous sections, I have discussed the liberal notion of reflective autonomy as part of the more general structure of the liberal notion of responsible agency. It assumes, as I have tried to show, that (a) action presupposes *volitional* autonomy and (b) reflective volitional autonomy in its turn presupposes *epistemic* autonomy. Both notions of autonomy require that we understand human subjectivity as having a *reflective*, more specifically *hierarchical*, structure. In the end, this allows us to understand actions as originating in *reflective epistemic autonomy*, in the sense of knowing one's will and one's beliefs (at least those beliefs that are relevant for the formation of one's will). In the previous sections, I have assumed rather than explained that and how the reflective structure of human subjectivity should, or can, be understood within a mentalist framework. In this section, I will explain the outlines of the mentalist framework and show how reflective epistemic autonomy may be understood within this framework as *introspective* self-knowledge. In the next chapter, I will discuss several arguments that aim to undermine both reflective volitional autonomy and reflective epistemic autonomy. But the force of these arguments can be understood only on the assumption that the liberal notion of autonomy presupposes an idea of self-knowledge (in a broad sense, including, for example, proprioceptive knowing what one's actions are, but also what one's intentions for actions are) that is described as basically a *cognitive-psychological* process: as *introspection*.⁸⁸

We have seen that the notion of reflective epistemic autonomy is already embedded within a vocabulary of consciousness. This does not necessarily hold for the notion of reflective volitional autonomy. This is a peculiarity of debates on action and free will more generally, as opposed to debates on reflective belief. Below I will comment on this and suggest that we should nevertheless understand the debate on action and free will against a mentalist background. Next, I will explain the mentalist framework by discussing what I think are some of its basic theses. As it provides the background for many, various discussions, I will focus on those theses that have shaped the notion of introspective self-knowledge. My discussion will not be detailed. In the next chapter, I will provide more detail when I discuss various 'mentalist' arguments against reflective volitional autonomy and reflective epistemic autonomy.

As I have suggested earlier (chapter 1, section 4), mentalism is not so much a specific theory but a general, basic *framework*. This explains, in part, the difficulties of making it explicit. Theorists working *within* this framework will usually evoke only the notions they disagree about, not the framework that functions as the tacitly agreed background for their discussions. In many works of philosophy of mind, therefore, this framework

88 Since the end of the nineteenth century, 'introspection' has become the technical term in debates at the intersection between psychology and philosophy of mind. It is closely connected to those other technical terms 'reflection' and 'consciousness' (or 'self-consciousness').

cannot be found publicly on the stage, even if it is peeping through the curtains. It sometimes takes critical outsiders to make the effort of articulating the framework.⁸⁹ Another difficulty is that, while the terms ‘reflection’ and ‘consciousness’ can be taken to belong to different conceptual histories, they also belong to our ordinary language. Not every philosopher’s use of a term is a technical use. Mapping the terminology of reflection onto the terminology of consciousness is therefore risky. Not every political philosopher stressing the importance of the reflective capacities can therefore be interpreted as being committed to a mentalist framework. Similar difficulties arise for the interpretation of analyses of action and free will. It may be conjectured that for notions of human agency, theorists have also employed the mentalist vocabulary. And yet, this is not the case. In a general discussion of notions of agency, in the context of debates on free will, O’Connor observes:

‘It is a remarkable feature of most accounts of free will that they give no essential role to conscious awareness. One has the impression that an intelligent automata [sic] could conceivably satisfy the conditions set by these accounts – something that is very counterintuitive’ (2000:122).⁹⁰

That the vocabulary of consciousness is absent from debates on action and free will is mainly the result of distinct historical developments within philosophy. It has no theoretical significance, I think. Below I will briefly provide some historical background to my suggestion. Next, I will suggest how we may understand Frankfurt’s account (1971) as the missing link between the determinism debate and the debate on introspective self-knowledge.

The new scientific worldview that came into being during the early modern period had similar philosophical implications for both the notions of human action and the mind. The problems these raised for each notion were nevertheless treated in relatively distinct domains of philosophy, each of which developed its own vocabulary to address these problems.

For analysing ‘mind’ as a philosophical problem, the vocabulary of consciousness was developed. In due time this was further developed, particularly within the Anglo-Saxon/-American tradition of philosophy, into a mentalist framework. In the early modern period, the term ‘conscious’ became a new, philosophical term in having the meaning of the mind’s self-knowledge (Davies 2008), at first of the mind as an epistemological self, later as a moral self (Tauber 2005). Descartes is generally understood as having introduced this new vocabulary. Philosophers of mind also understand him as having introduced the mentalist

89 By ‘outsiders’, I mean outsiders to either philosophy of mind or to the Anglo-American tradition of philosophy. See e.g. Schnädelbach (1977); Descombes (1995); Gethmann (2007). But also see Collins (1987).

90 Also, see Caruso (2012), who shares this observation.

idea of introspection.⁹¹ Various historical overviews indicate that Anglo-Saxon (later on: Anglo-American) philosophers favoured an interpretation of Descartes's *res cogitans* in terms of an 'empirical reflection', i.e. consciousness in a material sense (Nicholls and Liebscher 2010a).⁹² The understanding of consciousness in an empirical, material sense made it possible to offer proposals for understanding the mind in empiricist-materialist terms.⁹³ Consciousness was developed into a more specific notion as introspection. This was specific for the Anglo-Saxon/-American tradition of philosophy.⁹⁴ The later relation of philosophy of mind towards consciousness and introspection is a complex one. The notions of consciousness and introspection seemed to lose their relevance during the twentieth century.⁹⁵ In recent decades there has, however, been a renewed interest in these notions.⁹⁶ What complicates the history of the vocabulary of consciousness is that originally it was not linked to any notion of the 'unconscious'. Only later, starting from the nineteenth century, did it enter the vocabulary of consciousness. In the next chapter (section 2), I will make more comments on this. The vocabulary of consciousness does not necessarily function as a naturalist meta-language, i.e. as a mentalist framework. But it has nevertheless developed into a framework that naturalists use to understand *consciousness* – and with it all the notions that presuppose consciousness, such as rationality, deliberation – according to a mechanistic picture of the world.

For analysing 'human action', a different vocabulary was developed in a distinct but parallel debate. It took 'free will' rather than 'consciousness' as its central notion, but it is in a sense the equivalent of 'conscious state'. The historical background for the debate on action and free will be explained in more detail in chapter 5 (section 2). For now, it suffices to say that current discussions on moral responsibility, which connect human

91 For a critical discussion of this view: Hatfield 2011.

92 'Empirical reflection' is not a common phrase in philosophical vocabulary. But see Gasché (1986); Sandywell (1996); Swindal (1999). I found this phrase in German ('empirische Reflexion') and French ('réflexion empirique') philosophical literature mainly connected to discussions on Locke, Kant and Fichte. More generally on reflection: see Sandywell (1996). For tracing the beginning of the Anglo-Saxon tradition of 'empirical reflection', Locke is often cited: 'Consciousness is the perception of what passes in a Man's own mind' (1690/1975, II, 1, 9). For a critical discussion of this view: see Rabb 1979.

93 Not all Anglo-Saxon philosophers agreed with choosing this strategy. Thomas Reid and other Scottish philosophers argued against Humean empiricism and associationism. Later, British idealists, influenced by the German idealists, also took another road. But both counter-movements largely lost ground, within the Anglo-American context at least, to what would develop into analytical philosophy and philosophy of mind, partly following the footsteps of German psychologists, who themselves had rebelled against German idealists: see Richards (1992).

94 See Boring (1953); Danziger (1980) and (2001); Brock (2013). Nicholls and Liebscher (2010a), but also Richards (1992), suggest an alternative German tradition that took a functional approach towards consciousness and understood the mind as holistic, active and transcendental.

95 On the decline in interest for consciousness: see Manson (2002); Crane (2019). On the decline in interest for introspection: see Lyons (1986). For critical views on the role of introspection in the history of psychology: Leahey (1992); Costall (2006).

96 See e.g.: Siewert (1998); Smithies and Stoljar (2012); Kriegel (2014); Crane (2019). For a renewed interest in introspection within psychology: see Overgaard (2006).

action and determinism, can be traced to early modern debates on human action.⁹⁷

We have already seen that Frankfurt's article has provided a model for notions of liberal autonomy (section 3). Frankfurt's account of free will does not rely, however, on mentalist terminology: the adjective 'conscious' and the 'unconscious' both appear only once.⁹⁸ His account has nevertheless been noticed by philosophers of mind who do use

the vocabulary of consciousness. Although King and Carruthers observe that Frankfurt does not use this vocabulary, they nevertheless recognise in the structure of his argument an analysis that can easily be assimilated into a mentalist vocabulary:⁹⁹

'We submit that although neither Frankfurt nor Watson uses the language of consciousness, the reflections, higher-order motives, and value judgments that characterize an agent's real self (on such views) must be conscious ones' (2012:219).

This still does not show how Frankfurt's account of free will, as an account that is recast in the vocabulary of consciousness, may be linked to the debate on action and free will. But here, Frankfurt's article also functions as the missing link. As theorists developed their notions on autonomy, modelling them on the hierarchical structure that Frankfurt had introduced, theorists who worked in parallel discussions on free will were looking for similar ways to theorise agency. For some time, these parallels seemed to go unnoticed until Fischer commented:

'For many years I have been struck by the fact that there are "parallel literatures" which discuss "moral responsibility" and "autonomy." In many ways the literatures are isomorphic. For example, "hierarchical" approaches (involving the apparatus of higher-order mental states – states directed at [say] "first-order" mental states) are employed in the literature on moral responsibility as well as on autonomy' (1999:98).¹⁰⁰

97 See McKeon (1957). The overview McKeon offers of his etymological findings on responsibility (see 1957:67-68) is still the most instructive and comprehensive. For example, Henriot (1977) does not really add, I think, any new, conclusive etymological insights.

98 Frankfurt: 'The desires in question may be conscious or unconscious' (1971:8).

99 Frankfurt may be more indebted to the mentalist framework than King and Carruthers seem to recognise. As he makes clear in the opening paragraph of his article (1971), Frankfurt responds critically to an argument in which Strawson tries to understand the notion of personhood in terms of ascribing both corporeal characteristics and consciousness. For his own notion of personhood, Frankfurt chooses to shift from Strawson's ascriptive (third-person) approach towards a first-person approach. This may indicate a disagreement with Strawson's ascriptive approach itself.

100 Also, see Oshana (2002) and Arpaly (2003, esp. chapter 4) and (2004).

In recent decades, Frankfurt's account of freedom – 'Freedom of the Will and the Concept of a Person' – has exerted much influence in naturalist debates on moral responsibility. We will take a closer look at this in the next chapter (sections 2 and 3). This is no coincidence as both debates largely deal with similar issues to which Frankfurt's article had equally much to contribute.

Now I will turn to some explanation of the mentalist framework. A more familiar or more conventional term for mentalism is perhaps cognitivism. Although it is closely related to mentalism, the disadvantage of this term is that it is too closely tied to cognitive psychology. The notion of cognitivism involves conceptual strategies that certain philosophers would perhaps reject, even if, on my account, they would be committed to a mentalist framework. Mentalism, I would suggest, should be described at an even more abstract level. Below I will attempt to articulate what mentalism is by analysing it into various theses.

The Vehicle Thesis. From the natural scientific perspective, the world consists only of matter. It seems to resist granting any place to our thoughts. And yet, it seems they must have someplace in this world. But how? Of course, this constitutes the body-mind problem.¹⁰¹ But solving this problem depends on how the mind is understood. What is distinctive about our thoughts, our mind, is what philosophers nowadays sometimes call 'intentionality'.¹⁰² This technical term is somewhat unfortunate in English, referring to the 'meaning' of our thoughts rather than, in ordinary English, to what we are planning to do.¹⁰³ Nevertheless, the intentionality of our thoughts is assumed to comprise, next to, for example, our beliefs, feelings, also our intentions (in the ordinary sense). According to mentalism, whenever any of these forms of intentionality can be attributed to a person, for example, a belief that people should be treated equally, this belief exists in that person *as* a mental state. Brandom (2004b) has called this the *vehicled approach*. Mental states function as the 'vehicles' for people's intentionality. Talk about intentionality as mental states, implying that people's beliefs exist in their heads as inner episodes, allows for the possibility, theoretically at least, to describe people's intentionality also in non-intentional terms. A belief that may be attributed to some person may therefore be described in intentional terms – 'people should be treated equally' – and in non-intentional terms, for example, in terms of brain activity.

101 Cf. Matson (1966) and Peter King (2007), who both suggest that the discussion on mind-body is distinctively modern.

102 Cf. Kitchener: 'The problem of meaning is (arguably) the central problem in 20th century philosophy. Whether it takes the linguistic form of an inquiry into semantics, a mentalistic form of an inquiry into intentionality, or a cultural-historical form of an inquiry into the meaning of a text or an action, the problem of meaning has been at the center of philosophical discussion both in Anglo-Saxon analytic philosophy and in European phenomenology, structuralism, hermeneutics, and poststructuralism' (1994:279).

103 Searle (1979) introduced the proposal to distinguish 'intentionality' in the philosophical sense from 'intentionality' in the ordinary sense by capitalising the first term. I will not follow his suggestion, though.

The Consciousness Thesis. According to mentalism, mental states are either conscious or unconscious.¹⁰⁴ This thesis is needed as a first step to theorising introspective self-knowledge. The conscious/unconscious distinction has been interpreted in various ways, but its basically dichotomous way of dividing mental states into two ‘types’ has for many philosophers of mind worked well. It made it possible, for example, as we will see in the next chapter, to account for the phenomenon that people sometimes act in a different way than what they explicitly believe.

The Searchlight Thesis. The question of how to define ‘consciousness’ has provoked much discussion. Consciousness may mean, for example, how it feels like to be in a certain mental state. At this point, philosophers of mind will usually mention the example of seeing redness (for some reason, seeing ‘whiteness’ or ‘blackness’ is not interesting). In such cases, consciousness is defined on what it by its own. Another possibility is to define a mental state in terms of its relationship with another mental state. A recurring idea has been that mental states are able to function as a kind of searchlight towards other mental states.¹⁰⁵ In the last decades, various theorists have proposed that these mental states are unconscious: ‘What makes a mental state conscious (illuminated) is the fact that it is taken as an object by a relevant higher-order state’ (Zahavi 2007:268).¹⁰⁶ A point of dissension, however, is whether the searchlight mental state is itself a conscious state or whether it makes sense to understand ‘consciousness’ in terms of such relations.¹⁰⁷ Defining ‘consciousness’ – or more accurate perhaps: mental states – in terms of their relations nevertheless provides the first step to understand the self-directedness of the mind.¹⁰⁸

The First-Person Singular Thesis. Distinctive of (conscious) mental states is a first-person singular perspective. Mental states represent a first-person singular perspective. Combined with the previous thesis, it may explain how people, introspectively, recognise beliefs as their own.

The Introspection Thesis. The previous theses provide the ingredients for a mentalist notion of introspection. A person may learn about the world by observing: by looking, feeling, hearing. She may also learn about herself, for example, by interviewing the people that know her best, her parents, her friends. But she may also learn about the world and herself by becoming conscious of her own mental states. This process is called

104 Claiming that a mental state is conscious should be confused with claiming that a person is conscious.

105 I borrow this metaphor from Sher (2009), who uses the expression ‘searchlight view’ to distinguish the standard moral view that ‘an agent’s responsibility extends only as far as his awareness of what he is doing’ (2009:4). Sher connects the searchlight view explicitly to a type of moral psychology in which ‘consciousness’ plays a distinctive role to pick out the beliefs and decisions that have significance for our responsibility.

106 Their theories are called ‘higher-order’ theories. See e.g. Gennaro (2004); Kriegel and Williford (2006). For an overview: Carruthers (2016).

107 For criticisms: see, e.g. Siewert (1998); Zahavi (2007).

108 See, e.g. Smithies and Stoljar: ‘A mental state is conscious if and only if one knows by introspection that one is in that mental state’ (2012:20).

‘introspection’: ‘Introspection, as the term is used in contemporary philosophy of mind, is a means of learning about one’s own currently ongoing, or perhaps very recently past, mental states or processes’ (Schwitzgebel 2016:1).¹⁰⁹ The previous theses contribute in various ways to the notion of introspection. Introspecting, therefore, is one way of acquiring knowledge about the world and oneself.

The Difference Thesis. The way that people acquire knowledge through introspection is different from other ways of knowing, such as perception: ‘We might form the belief that someone else is happy on the basis of perception – for example, by perceiving her behavior. But a person typically does not have to observe her own behavior in order to determine whether she is happy. Rather, one makes this determination by introspecting’ (Kind 2018). What is distinctive about introspection may be thought of as either a psychological or cognitive process or even a cognitive-psychological process.¹¹⁰

Theorists on introspection will probably not endorse all the above theses. At least they disagree on how to understand the theses. One of the dividing lines among theorists on introspection turns on whether they claim that aspects of mentality are real or just illusionary. Some claim that mental processes involve, for example, a real self or a real first-person singular perspective and therefore are likely to emphasise introspection as a *cognitive* process.¹¹¹ Others will deny that these aspects present anything real. These theorists will agree that people, in fact, *experience* their thoughts as involving a special kind of knowledge (introspection) and as involving a real self, or even as experiencing a form of introspection (in the sense of a conscious process). But different from the first group of theorists, they will claim that these experiences are illusionary. They will therefore emphasise that mental processes are psychological. They nevertheless accept the basic conceptual structures of the mentalist framework.

The Hierarchy Thesis (or: the Control Thesis). Frankfurt’s account of freedom of the will in terms of first- and second-order desires/wants can be understood as fitting within the mentalist structure of introspection. Frankfurt’s account cannot, however, be sufficiently understood based on only the theses already mentioned. What he distinguishes as first- and second-order desires/wants seem to be mental states that the person is conscious of. Although the *relationship* between these mental states might be analysed in terms of a searchlight function, what Frankfurt is interested in is really a different kind of relation: a hierarchical relation. What is still absent in the previous theses is a thesis stating that mental states may be understood in relation to other mental states in terms of control.¹¹² I will call it the Hierarchy (or: Control) Thesis. For Frankfurt, this makes sense in his analyses. But I think this thesis is also part of mentalism. The discussion between

109 For other similar definitions of introspection: see Rosenthal (1999); Kind (2018);

110 For a discussion: see Smithies and Stoljar (2012).

111 See Moran (2001); Baker (2013).

112 Another way of saying that mental states can have a control function would perhaps be that they are goal-directed, roughly in the sense that Bargh (1990) discusses.

theorists on introspection will centre, once again, on whether the sense of control is real or only illusory. Theorists may favour a purely psychological approach. They may nevertheless be understood as committed to the Hierarchy Thesis if they claim that the relevant mental states exerting control on other mental states are unconscious rather than conscious. Theorists who favour a cognitive approach will claim that the relevant mental states are conscious and that their control is real. They may disagree, however, on how to qualify the nature of this control: as either an act of will (voluntaristic view) or the force of reason (reasons-responsiveness view).¹¹³

2.6. Conclusion

In this chapter, I have identified and discussed two central features of the mentalist notion of responsibility: authorising and introspective self-knowledge. Nowadays, one of the dominant views on ‘authorising’ is to understand it in terms of autonomy, i.e. as grounding authorising in one’s real self. To spell out how autonomy works, and therefore to explain the cognitive-psychological conditions of responsibility, liberal theorists rely on a notion of introspective self-knowledge. It presupposes a mentalist view on the human mind to which the conscious-unconscious distinction is central. I suggested a reconstruction of two notions that show how introspective self-knowledge is needed for autonomy: reflective volitional autonomy and reflective epistemic autonomy.

I have highlighted these two features because they provide the two main ingredients for understanding epistemic autonomy. While I think that authorising should be kept as a feature of the liberal notion of responsibility, the problem is first and foremost with understanding self-knowledge as introspective. The mentalist notion assumes that this feature provides a solution to the problem of agential self-ignorance, which I discussed by homing in on manipulation as a case of (causing) agential self-ignorance. Whether this is the case will be explored in chapter 4. First, I will introduce and discuss, in the next chapter, yet another case of agential self-ignorance: hidden prejudices.

¹¹³ In chapter 5, section 4, I will say more about this.

3

Chapter 3

The veil
of prejudice

3.1. Introduction

This chapter aims to introduce and explain the veil of prejudice that hides prejudices from people's self-knowledge. I will first explain it in terms of 'implicit bias', a notion that social psychologists use for their research into prejudice.¹ In the following chapter, it allows me to explore whether people are responsible for their prejudices and, as we will see, whether they are responsible for any action at all. A mentalist framework supports the liberal notion of responsibility, as we have seen, and the notion of implicit bias, as we will see. In later chapters, I will be critical about this framework and, therefore, about the notions it supports, such as implicit bias. In this chapter, I will also be critical about the notion of implicit bias: not for its mentalist framework, which I will accept for now, but to improve it. The notion of 'implicit bias' provides an interesting proposal to understand 'hidden prejudices' in psychological terms, but it can still be improved by extending the type of practices it is supposed to explain. In a critical discussion of views by Fricker and Bourdieu, I will propose that we include, in our understanding of the veil of prejudice, not only *practices of discrimination* but also *practices of silencing*. First, I will, in this introduction, explain what will and will not be part of how I understand this notion. Next, I will show how I will set up this chapter.

My discussion will focus on two questions: (1) how to qualify 'hidden' prejudices people have but are ignorant about? (2) What explains practices of prejudices as a *social phenomenon*? In other words: what may explain why practices of prejudices persist? To answer the first question, I will introduce and explain the notion of implicit bias (section 1). In the sections that follow, I will deal with the second question. Here I will take my cue not from the research on implicit bias but from Fricker's analysis of *epistemic injustice*. Fricker does not present it in terms of the liberal-mentalist notion of responsibility, nor of implicit bias. But it nevertheless indicates a commitment to an epistemic-mentalist approach to prejudices that also characterises the liberal-mentalist notion of responsibility and the notion of implicit bias.² While research on implicit bias takes for granted that hidden prejudice (implicit bias) is linked to practices of discrimination, Fricker's analysis has the advantage of trying to get beyond this link. But I do not think she succeeds. I will first briefly explain her analysis of epistemic injustice (section 3) and then critically discuss it (section 4). Next, I will turn to Bourdieu and draw on some of his notions and insights, his concept of *habitus*, of course, but also his analysis on censorship (section 5). His analysis allows me to show that prejudiced people perform not only *practices of discrimination* but also, distinct from these, *practices of silencing* that help to preserve prejudice in society (section 6).

1 For an overview of the philosophical issues: Brownstein (2019).

2 For Fricker's commitment to a mentalist vocabulary: see this chapter, note 45.

3.2. Implicit bias: a mentalist notion of hidden prejudice

In this section, I will introduce the social-psychological version of hidden prejudice or, as social psychologists call it: implicit bias. I will briefly explain the context, how social psychologists understand ‘implicit bias’, what kind of discriminatory practices it motivates people to do and how it is measured. (When I talk about ‘implicit bias’ in this and the next chapters, I mean to refer only to what social psychologists mean by it. My own phrase ‘hidden prejudice’ is intended as an umbrella term for specific notions of hidden prejudices.³)

Prejudice has been an important philosophical notion during the Enlightenment period. Part of the heritage of ‘prejudice’ as an Enlightenment concept is the insight that people suffer from *ignorance* because, while knowing they have certain beliefs, they are ignorant about the false grounds their beliefs are based on.⁴ The use of this concept was at first limited to issues of epistemology, i.e. to beliefs people have about the world. Later, it was developed into a moral concept. In the nineteenth and twentieth century, philosophers lost their interest in discussions that were specifically tied to the term ‘prejudice’ (*praejudicium*, *Vorurteil*, *préjugé*). While it had acquired to some extent a critical role in epistemological and moral analyses, some of its critical functions were taken over by other notions, for example, by the term and notion ‘ideology’, a Marxian concept that is congenial to the Enlightenment concept of prejudice, but whose conceptual roots cannot be traced back to it.⁵ Meanwhile, the concept of prejudice had been taken up, in the Anglo-American context, mainly by sociologists and social psychologists, mostly taking

3 Much of the research on implicit bias has honed in on *racial* stereotypes, as we have seen, probably because the first generation of social psychologists on implicit bias belong to American society. Although discrimination of, for example, Roma people in Europe has become a political topic in recent decades, social-psychological on implicit bias against Roma people seems rare. But see Růžička (2014); Lanting et al. (2018, unfortunately only an abstract). Also, see Marek et al. (2020), which investigates hidden prejudices of healthcare staff against Roma patients (while the researchers use the term ‘implicit bias’, their research does not involve any implicit bias test). For a more general discussion on discrimination in Europe, especially focussing on discrimination against Romani: Tileagă (2016). It also includes a discussion of implicit bias, but does not contain any specific research on implicit bias against Romani.

4 See Schneiders (1983); Beetz (1983); Dorschel (2001, chapter 1); Godel (2007); Reisinger and Scholz (2019). The concept of prejudice even played a role in European travel literature during the Enlightenment period: Kagel (2010). Schneiders (1983) observes that English and French analyses were mostly focused on specific cases of prejudice, while German analyses developed during the Enlightenment into more general, systematic theories of prejudice. Hundert (1987) explains how ‘prejudice’ was developed, as a *critical* notion, in opposition to the notion of ‘common sense’.

5 For the critical role of ‘prejudice’ in political analysis: see Schneiders (1983, esp. chapter V). For a brief discussion of the conceptual relations between ‘prejudice’ and ‘ideology’: see Schneiders (1983, chapter I). For a brief historical overview of the concept ‘ideology’: see Thompson (1990, chapter 1).

racism in the United States as their prime example of prejudice.⁶ But the various notions of prejudice they forged, despite using the same term as Enlightenment philosophers, seem to have no direct link with the Enlightenment notion of prejudice. The notion of prejudice, so it seems, was reinvented to address what was increasingly recognised, during the twentieth century, as a pressing, moral-political problem in western societies: racist prejudice (against Jews and African Americans).⁷ By now, research into prejudice and discrimination is rapidly growing.⁸

Within the Anglo-American context, the notion of prejudice changed during the second half of the twentieth century from old to new prejudice, from overt to covert prejudice.⁹ For a long time, Allport's seminal work *The Nature of Prejudice* (1954) has dominated social-psychological research on prejudice. Ever since, social psychologists gradually began to take an interest in the divergence of what their respondents said and how they behaved. For example, they had noticed that 'self-report questionnaires did not always capture people's true attitudes toward members of racial outgroups' (Amodio and Mendoza 2010:354). Such observations were reinforced by findings from various experiments outside the field of research on prejudice and discrimination, for example, in

-
- 6 In general, the experiences of the Second World War contributed to an increasing interest in performing research into prejudice and discrimination (Dovidio 2001; Reicher 2012). Before the war, sociological research into practices of racist prejudice had already started with Du Bois. Virtually all his writings, as soon as his (1898), deals with it. A new line of sociological research started in the mid-1960s with research into racist prejudices: see Carmichael and Hamilton (1967); Pincus (1999). Psychological research started in the 1920s. Important publications in the 1950s: Adorno et al. (1950); Saenger (1953); Allport (1954). For a historical overview of the social scientific study of prejudice: Stroebe and Insko (1989); Duckitt (2010). Interestingly Du Bois seems to have had little or no influence on concepts of prejudice in social psychology. He is not mentioned by Allport [1954] or by Duckitt (2010). In the 1990s Gaines and Reed (1994) and (1995) have argued for his relevance to theorising in social psychology. Outside the Anglo-American context, there was also a renewed interest in research into prejudice, as in Germany, for example, where it was largely initiated by the 'Frankfurt School' (see, also for references, Schneiders 1983, first chapter).
- 7 During the twentieth century, social psychologists struggled to theorise prejudice. The core concept from which they started was the 'stereotype', introduced as a socio-political metaphor by Lippman (1922). At first, they were tempted to identify prejudices by making generalisations about other people. But this proved unfruitful, as they acknowledged that this is really a general feature of people's cognition. Therefore, a second step was to distinguish prejudice from generalisations (stereotypes) in defining it as an attitude (in a social psychological sense, not to be confused by philosophical uses of 'attitude'; see Mandelbaum 2014). Even then, the notion of prejudice retains a link with stereotypes because it includes, as one of its elements, a stereotype about people from other social groups. For a more detailed discussion: see Stroebe and Insko (1989). The notion of 'stereotype' is unknown to Enlightenment theories of prejudice. I have found no specific references in social-psychological literature to early modern analyses of prejudice. Allport starts his (1954) with the briefest possible summary of the history of the term 'prejudice', one of the meanings being 'prejudgement'. His main source, mentioned by himself, is Murray (1909).
- 8 See Dovidio et al. (2010). Also, see Beeghly and Madva (2020).
- 9 See Brown (2010, chapter 7); Pehrson and Leach (2012). For an argument that the continuity between old and new prejudice is probably greater than is usually assumed: see Leach (2005). The continuity between the old and new prejudice, both being explored within social psychology, is nevertheless probably greater than between the Enlightenment notion of 'prejudice' and the social-psychological notion of 'prejudice' in the twentieth century.

the classic study by Nisbett and Wilson (1977), indicating that people may have limited introspective access to their own choices, judgements and behaviour. For that reason, they began to look for ‘experiments involving hidden manipulations and unobtrusive measures of racism’ (Crosby, Bromley and Saxe 1980:546). This also meant a shift in what practices were studied. While Allport’s study of prejudice as antipathy was still mainly focussed on *overt* practices of prejudice, such as exclusion and violence (Dovidio et al. 2005), the new experiments instead attended to *covert* practices of prejudice. They involved, for example, a black or a white person requesting a dime for cookies or milk or a driver (either white or black) who stood by his/her car that appeared to have broken down.¹⁰

The shift from old to new prejudice coincided with, and perhaps was partly informed by, a transition in public expression and representation from overt to covert racism. While it may be claimed that certain prejudices, for example, racism in the United States, had always been overt, explicit in law and legal decisions, but also in public debate, research suggests that prejudice (particularly racism) has ‘gone underground’, becoming less overt, not just in the United States, but also in Western Europe.¹¹ One context that provides possible clues to understanding was that during the second half of the twentieth century, civil rights gained wider attention in Western societies, which resulted in more egalitarian laws and policies. In one interpretation of this development, people began to conceal their racist views.¹²

Social psychologists struggled for some time to find an appropriate theoretical framework for the new, covert prejudice. An important step was taken by Devine, who proposed – against the then-prevailing view among view social psychologists that having prejudices was the inevitable result of any person’s ‘social heritage of a society’ – that a distinction should be made between ‘knowledge of a cultural stereotype and acceptance or endorsement of the stereotype’ (1989:5). She argued, therefore, ‘that stereotypes and personal beliefs are conceptually distinct cognitive structures’ (1989:5). This implied, for example, that having knowledge of a stereotype does not necessarily imply agreeing to that stereotype. It also implied, crucial for research on implicit bias, that people may behave in a way that agrees with (prejudicial) stereotypes they have knowledge of but which does not agree with their explicit, avowed beliefs.

10 These examples are taken from the overview that Crosby, Bromley and Saxe provide (2008:550–552, Table 1).

11 For western Europe: Pettigrew and Meertens (1995). For the United States: one of the first to notice this ‘altered nature of racial discrimination, from blatantly exclusionary practices to more subtle, procedural, ostensibly “non-racial” forms centred upon demographic trends, housing patterns, and spatial arrangements’ has been Pettigrew (1979:114). The racist variety of the veil of prejudice has been given different names by philosophers, sociologists and social psychologists (in most cases to capture the situation in the United States): ‘liberal racism’ (Sullivan 2006), ‘silent racism’ (Trepagnier 2010), ‘modern racism’ (McConahay, Hardee and Batts 1981), ‘new racism’ (Sniderman, Piazza, Tetlock and Kendrick 1991); ‘aversive racism’ (Dovidio and Gaertner 2004). For still other varieties: see Pettigrew and Meertens (1995) and Bonilla-Silva, Lewis and Embrick (2004).

12 For this interpretation: see Crosby, Bromley and Saxe (1980).

Devine's proposal required a new theoretical frame.¹³ It had to be able to make sense of prejudiced behaviour while their agents disapproved the explicit prejudices that, in some hidden way, informed the prejudiced behaviour. How was that to be done? One possible notion was the 'unconscious'. It was not, of course, an unfamiliar notion for psychologists. During the twentieth century, Freud had become the leading theorist of this notion.¹⁴ In recent decades various scholars have used the Freudian 'unconscious' for their explanations of hidden prejudice.¹⁵ Social psychologists working on implicit bias, however, rejected it for its empirical elusiveness: 'Freud's views of unconscious mechanisms were embedded in a theory that never achieved conclusive support among scientists, despite many empirical theory-testing efforts in the middle third of the twentieth century. Consequently, most psychologists have abandoned Freud's psychoanalytical theory of unconscious mental processes' (Greenwald and Krieger 2006:945).¹⁶ An alternative theoretical frame was found within the overall *cognitive* approach that had begun to dominate psychology.¹⁷ Next, within this approach, the required conceptual resources were derived from various domains of psychological research. Research in one domain was focused on how people learn and store concepts from categories that are semantically related. It gave insight into the automaticity of how people process social information. It showed, for example, that whenever different words had meanings that were closely related, people were able to recognise such words faster.¹⁸ People were assumed to have no control over the speed (the 'automaticity') of their reaction time. Such insights were helpful in designing experiments for detecting prejudice, for example, by testing how fast respondents would categorise stereotype words more quickly with either 'NEGRO' or 'WHITE'.¹⁹ Yet another domain of research was devoted to the functioning of memory processes. Research findings suggested two different systems

13 See Amodio and Mendoza (2010).

14 Allport has not systematically analysed prejudices in terms of 'unconscious processes', probably because his focus on overt practices did not require this, of course, but also, as Dovidio, Glick and Rudman (2005:12) suggest, 'because of his unwillingness to fully embrace psychoanalysis'.

15 See Kovel (1970); Lawrence III (1987); Young-Bruehl (1996); Mills and Polanowski (1997); Altman and Tiemann (2004); Sullivan (2006). More recently, an attempt has been made to reconcile the Freudian and cognitive unconscious. See, e.g. Talvitie (2009). Noteworthy is also an attempt to forge a notion of hidden (racial) prejudice based on Harold Garfinkel's *Studies in Ethnomethodology* (1967): see López (2000).

16 Interestingly, in responding to new notions of 'unconscious' (as compared to the Freudian 'unconscious) in cognitive neuroscience and cognitive psychology, Woody and Phillips (1995) argue that the Freudian 'unconscious' is better understood as an essentially 'hermeneutic concept' rather than an empirical concept (which is how Greenwald and Krieger understand the Freudian 'unconscious'). Stroebe and Insko (1989), in their turn, claim that the psychoanalytical frame disappeared, not because it was 'refuted empirically', but because it 'simply went out of fashion' (1989:3).

17 Stroebe and Insko (1989) observe how in successive editions of the *Handbook of Social Psychology* (first edition in 1954, second edition in 1969; third edition in 1985), the number of approaches towards prejudice and discrimination (initially comprising psychoanalytic, personality-oriented and still other approaches) was in the end reduced to only one: the cognitive approach.

18 See Meyer and Schvaneveldt (1976).

19 See Gaertner and McLaughlin (1983).

for memory, one for implicit memory processes, another for explicit memory processes. These and other research domains were integrated into one theoretical frame for research on implicit bias, to which the notion of *implicit social cognition* became central. Research on memory processes has provided the adjective ‘implicit’ in coining the phrase ‘implicit bias’, but we should realise that this adjective conceals a complex fusion of various psychological notions from different types of research. In the next chapter (section 5), I will discuss in more detail how to understand the ‘implicit’ in *implicit bias*. For now, it suffices that we can take ‘implicit’ to mean ‘unconscious’ in a sense that researchers on implicit bias have generally concluded that implicit biases ‘cannot be directly inferred through introspective awareness’ (Amodio and Mendoza 2010:355).²⁰

The ambition of research on implicit bias is to predict discriminatory behaviour. As I already suggested, the new research on prejudice as implicit bias, as compared to previous research (Allport and others), is dealing with another type of discriminatory practices: covert instead of overt. *Covert* discriminatory behaviour can be found in a wide variety of social domains: in the workplace, employment, housing, credit market, health care, education, criminal justice system, consumer market.²¹ And what do discriminated people experience? In figure 2, I provide a brief overview of what research on implicit belief has already revealed.²²

<i>Social domain:</i>	<i>What discriminated people may experience (as compared to other, less, or non-discriminated people):</i>
Health care	<ul style="list-style-type: none"> - They are more likely to have their physicians display less willingness to treat them immediately (Drewniak et al. 2016). - They are more likely to be judged by health professionals as lazy, stupid and worthless (Schwartz et al. 2003).
Education	<ul style="list-style-type: none"> - They are more likely to have their teachers having lower expectancies about their school performance (Bergh et al. 2010). - Being equally smart, they are more likely to obtain lower school grades (example from the United States: see MacCann and Roberts 2013).
Criminal justice	<ul style="list-style-type: none"> - They are more likely to be shot at by police officers (Plant and Peruche 2005).
Employment	<ul style="list-style-type: none"> - When equally qualified, they are less likely to be offered job interview opportunities (Rooth 2007). - When equally qualified, they are less likely to be hired (Rudman and Glick 2001).

Figure 2: What discriminated people may experience.

20 Also see Greenwald and Krieger (2006); Nosek and Jeffrey (2008).

21 For this overview, including examples: see Pager and Shepherd (2008). The studies that are mentioned, for example, in Jost et al. (2009) all within these categories.

22 What also provides an interesting source for examples of practices of prejudices are what Rowe (1990 and 2008) has called *micro-inequities*. Brennan (2013) recognises that micro-inequities may be understood as being motivated by implicit bias.

The prediction of discriminatory behaviour requires, ideally, the combination of two types of tests, one testing for implicit (prejudiced) attitudes, another testing for actual discriminatory behaviour. Relating the results should tell whether, and if so, to what extent, testing for implicit attitude predicts discriminatory behaviour. As my concern is mainly with explaining ‘implicit bias’, I will focus here only on methods for measuring ‘implicit bias’. Social psychologists have various methods of testing for implicit bias. The most widely used is probably the Implicit Association Test, also known as the IAT-test.²³ The IAT-test is accessible to anyone on the internet: <https://implicit.harvard.edu/implicit/takeatest.html>. It is an initiative of *Project Implicit*, founded by Greenwald, Banaji and Nosek, who are known for their research on implicit bias. This IAT test provides the opportunity for testing prejudice on various grounds, such as gender, disability, religion, race and weight. Here I will describe the test for racial prejudice. The core of the test consists of a series of questions that will ask you to press either one of two keys, for example, the ‘E’ or the ‘I’ key. Next, you will see, for each question, a photo in the middle of your screen: either an African American or a European American. Or you will see, again in the middle of your screen, a word that expresses either ‘bad’ (Selfish, Failure, Horrific, Gross, Sickening, Bothersome, Pain) or ‘good’ (Excellent, Enjoy, Friendship, Happy, Cheer, Laughing, Delight, Fantastic). For each series of questions, you will see in the left corner above: *Press “E” for*, for example, *Bad or European Americans*, and in the right corner: *Press “I” for*, for example, *Good or African Americans*. For each series of questions, either ‘Good’ and ‘Bad’ or ‘African Americans’ and ‘European Americans’ may change position. Any combination is possible. For each photo (for example, a photo of an African American) you see, or for each word (for example, Friendship) you see, you must press the right key as fast as possible. Suppose you must press “I” for *Good or African Americans*, then in the example below (figure 3), you would have to press “I”.

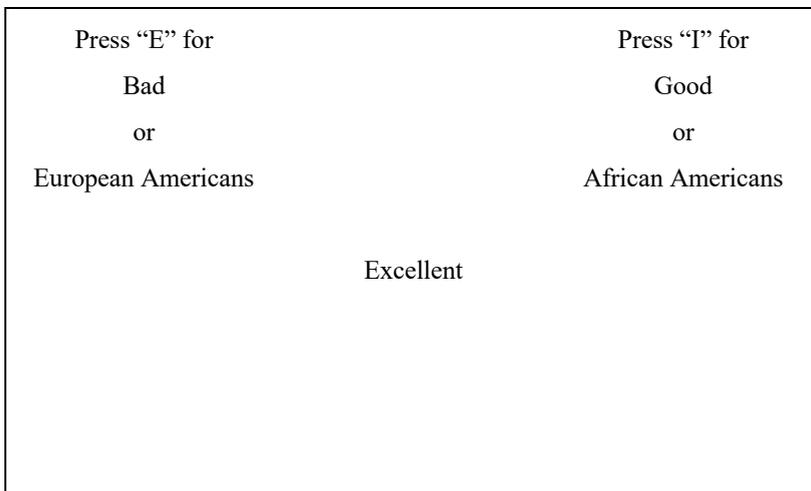


Figure 3: What a screen of an IAT-test may look like.

23 Introduced and described in: Greenwald, McGhee and Schwartz (1998).

If your response, when you had to press the same key for both *Good* and *European Americans*, was faster than when you had to press the same key for *Good* and *African Americans*, your result is described as ‘Automatic preference for European Americans over African Americans’. This will imply that you have an implicit bias against African Americans. Theoretically, the underlying idea is that you apparently have a stronger positive association with European Americans, i.e. an association that is better automated, less within the control of your reflective powers. The test is a reaction-time measure. People are asked to respond as quickly as possible.²⁴ The test is measuring latencies in responding. As the differences in responding to either Good + European Americans or Good + African Americans are too small to notice for respondents, the test is, for that reason, sometimes assumed to be resistant (but not immune) to faking.²⁵

From the previous discussion, it should be clear that social psychologists understand hidden prejudices as implicit biases: unconscious mental states that involve ‘fast, automatic, and difficult to control processes that encode stereotypes and evaluative content, and influence how we think and behave’ (Holroyd and Puddifoot (2020:313)). How exactly to conceptualise the content of such unconscious mental states is a matter of dispute among philosophers and social psychologists on implicit bias. A common view is that they consist of automatic associations.²⁶ But others have argued that it consists of propositionally structured beliefs.²⁷ For my argument, it suffices that an implicit bias is an unconscious mental state that establishes a link, in whatever way (an associative, propositional, et cetera, link), between people of certain social groups and certain negative stereotypes.

3.4. Prejudice as ignorance

In this section, I will start with introducing an *individualist-epistemic approach* to analysing dissonant practices of prejudices, for which I will draw on Fricker (2007) and argue for the limitations of this approach. In the following sections, I will make a gradual shift to a *socio-pragmatic approach* which, I think, can provide a richer, more complex notion of prejudice as ignorance.

The standard approach, which dates to Enlightenment thinking, is to analyse discursive prejudice as an essentially individual, epistemic phenomenon. Characteristic for this approach is the definition of prejudice in terms of individual judgements

24 See Lane et al. (2007).

25 For making such a claim (in this case for a different but similar testing model): see, e.g. Fazio, Jackson, Dunton and Williams (1995). For a discussion of this claim, based on various research reports on testing such models for their resistance to faking: see Greenwald, Poehlman, Uhlmann and Mahzarin (2009).

26 For an overview of various proposals for theorising the contents of implicit bias: see Holroyd et al. (2017, Section 6); Brownstein (2019, Section 2); Baston (2020, Chapter 4, Section 2).

27 See Levy (2015); Mandelbaum (2016). Still, others have proposed that implicit belief consists in an ‘alief’ (see Gendler 2008) or ‘mental imagery’ (Nanay *forthcoming*).

(prejudice being a certain type of judgement) and the focus on its epistemic conditions (in terms of truth, correctness, rationality). If we understand prejudices as a social phenomenon, they are to be considered as (pre)judgements that (a) pertain to negative qualifications, i.e. judgements on what negative attributes can and should be ascribed to people from certain social groups, and (b) are the result of faulty reasoning (false, incorrect, irrational judgements).²⁸ Enlightenment theorists thought that people who have irrational judgements could be said to be ignorant of the irrationality of their judgements. Nowadays, the epistemic approach seems to assume that – behind the veil of prejudice – people are ignorant, not just of the irrationality of their judgement, but of this judgement itself: the prejudice itself is hidden.²⁹

Representative of the epistemic approach, I think, is the analysis that Fricker provides of prejudices in her *Epistemic Injustice* (2007). What is valuable about her book is that she draws attention to a distinct type of practice of prejudice that she calls ‘epistemic injustice’ and which consists in the misrecognition of people from certain social groups (based on, e.g. gender, ethnicity, et cetera) as *knowers*. Fricker is interested in the relation between speaker and hearer and how the hearer judges whether he thinks a speaker is trustworthy, especially when this speaker belongs to another social group. Her analysis focuses on credibility judgements and the impact that stereotypical judgements, especially prejudices, have on them. If prejudice affects the hearer’s credibility judgement of the speaker, the result is testimonial injustice. Credibility judgements are the outcome, not of some reasoning or argumentation, but of certain processes of perception in cases of interaction between individuals. Whenever people are aware of this outcome, a credibility judgement is called ‘belief’. But what really interests Fricker is the process of perception itself: ‘credibility judgements are perceptions shaped not only by belief but also by emotional responses and by contents in the social imagination’ (2007:93). Perception as a process is either virtuous (unprejudiced) or prejudiced.

Virtuous perception is the sensitivity to perceive in someone’s performance those features that are relevant to making a sound moral judgement. If a speaker is talking about *x*, a virtuous perceiver would be able to tell whether the speaker is trustworthy in talking about *x*. One way of understanding how the perceiver has gained knowledge of *x* is based on the inferentialist model, which assumes that the perceiver reconstructs an argument of which the conclusion is *x*. Instead, Fricker argues for a model of non-inferentialist judgement. Virtuous perceivers can judge someone’s trustworthiness by some default of uncritical receptivity, a process that is ‘spontaneous and unreflective; it involves no argumentation or inference’ (2007:72). Virtuous perception opens the

28 Literature on prejudice acknowledges that prejudice may also be positive. But discussions almost always focus on the prejudices that are negative. I believe the definition of prejudiced judgement should also include judgements on categorisation, i.e. judgements on whether or how people can and should be viewed as certain types of social groups. See, e.g. Tilly (1998); Yanow (2003).

29 As this type of ignorance comes closer to what in psychology is called ‘unconscious’, the distinction is often blurred.

possibility of people having prejudiced beliefs but nevertheless being able to perceive in a virtuous way. Fricker mentions the example of Huckleberry Finn, who shares the prejudiced beliefs of his contemporaries but nevertheless can perceive the runaway slave Jim in an unprejudiced way.³⁰

Prejudiced perception involves a distortion of people's basic capacity for virtuous social perception. This process is also unreflective but differs from the process of virtuous perception. Fricker applies the epistemic approach to the prejudice involved in prejudiced perception: she defines prejudice primarily as a stereotypical judgement.³¹ To the extent that people are aware of their stereotypical judgements or prejudices, Fricker calls them 'beliefs'. But stereotypes may also become part, as images, of people's social imagination.³² Fricker understands the social imagination as consisting of judgements that take the form of images.³³ The images are stereotypical and contain certain epistemic commitments to the social world.³⁴ As people grow up as part of a social group, images of the social imagination structure their social consciousness. As a result, these images become part of their thinking and perceiving. If then stereotypical judgements are incorrect, unreliable or negative, they too inform people's social perception by shaping their patterns of judgements and motivations. But they operate directly, i.e. without doxastic mediation, and by stealth, 'not unconsciously in any strict, psycho-analytical sense, but without any focused awareness and without [their] permission, as we might put it' (2007:39). Fricker calls this process 'residual internalisation'. Whenever a prejudiced

30 In my terminology, this presents a case of dissonant prejudice, but one in which the explicit belief is prejudiced, and the behaviour is not.

31 Also, see Saul, who remarks that Fricker 'seems to be focused on prejudices as propositions explicitly believed/ endorsed' (Saul 2017:236). I should mention that Fricker uses the more specific phrase 'negative identity prejudice'. I will use the shorthand phrase 'prejudice'.

32 Fricker chooses to use the less heavily theoretical notion 'social imagination' instead of the kindred notion of 'social imaginary', which she finds problematic because of its psychoanalytical roots (Epistemic Injustice, p. 38).

33 For her notion of 'social imagination', Fricker refers to the work of Cornelius Castoriadis. I take Fricker's 'social imagination' to be related to such concepts as 'culture' (Geertz 1973), 'imagined community' (Anderson 1983), 'myth' (Yanow 1992), and 'social imaginary' (Taylor 2004). 'Ideology' does not fit in here, I believe. Ideologies are usually understood as more or less coherent systems of ideas that function as the legitimacy of power relations (cf. Thompson 1990, chap. 1). As such, this concept, compared to the other concepts mentioned here, relates to a more theoretical, more explicit view a social group has of itself or of other social groups.

34 For interesting examples of what shape these images may take, literally, and how pervasive they are: see Patricia A. Turner (1994).

image effectively shapes people's social perception, this is called *residual prejudice*.³⁵

The effect of residual prejudices is persistent. Even when people change their prejudiced beliefs during their lifetime into genuinely unprejudiced beliefs, they may still perceive their social world in a discriminatory way simply because the residual forms of their old prejudice retain their influence. People will have trouble becoming aware of those residual prejudices as these 'operate beneath the radar of people's ordinary doxastic self-scrutiny' (2007:40). They cannot easily stop them from interfering with their perceptions and moral performances.

Fricker's analysis reveals, I think, the limits of the epistemic approach for addressing the veil of prejudice as a moral-political problem. In their epistemic approach to prejudice, moral philosophers such as Fricker, but also social psychologists, usually acknowledge the social nature of prejudices for their social origins. Therefore, they invoke concepts such as culture or the social imagination in which prejudices are rooted. People acquire their prejudices due to the social imagination they share with other people in certain social contexts: 'The collective social imagination inevitably contains all manner of stereotypes, and that is the social atmosphere in which hearers much confront their interlocutors', as Fricker puts it (2007:38). But explanations of this type only move the problem forward. How did residual prejudices, or how did the social imagination itself, come into someone's head? As the social imagination cannot exist without individuals who embody them, it remains unclear what it exactly means that residual prejudices or the social imagination contribute to the formation of prejudices.³⁶ And why does the social imagination, or at least its prejudiced features, continue to exist? Fricker's handling of these problems again shows her basically epistemic-mentalist approach: she analyses them in terms of what a person may know, or not know, about the social imagination or about her own behaviour. She suggests, for example, the possibility that a person 'comes to notice a certain dissonance between her [avowed, unprejudiced] beliefs and her perceptual judgements' (2007:38). This may then be the starting point for developing an 'enhanced self-awareness that helps limit the impact of the prejudicial

35 Fricker seems to be reluctant to describe 'residual prejudice' in terms of the 'unconscious'. Instead, she takes recourse to various metaphorical phrases (e.g. 'residual internalization', 'residual form', 'beneath the radar'). From various remarks, it becomes clear that she wants to avoid, for reasons she does not explain, the association of the adjective 'unconscious' with the psychoanalytic notion, for example, when she dismisses this adjective in a further explanation of residual prejudice: 'not unconsciously in any strict, psychoanalytic sense' (2007:39; also see 2007:38, note 9). Her text reveals how the conscious/unconscious distinction nevertheless plays an important role for the notion of 'residual prejudice': whenever it is explained (in terms that I think are equivalent to 'unconscious'), it is often immediately contrasted with 'consciousness'-terminology ('self-consciousness', 'enhanced self-awareness'). For example: once she has dismissed the term 'unconscious' in its 'strict, psychoanalytic sense', she immediately adds, as an alternative phrase: 'without any focused awareness'. In (2007) Fricker does not refer to 'implicit bias'. She does, however, in (2016a) and (2016b). These show that whatever reasons Fricker has, to object to the psychoanalytic 'unconscious', these do not obtain for the 'unconscious' that social-psychologists use to explain their notion of implicit bias: '[...] prejudice can operate unconsciously or, as we have now learned to say, at the level of "implicit bias"' (2016b:162).

36 Cf. Pierre Bourdieu's critique of the 'occasionalist illusion' (1977:81). Also see Ostrow (1981:282).

residue on her credibility judgements' (2007:38). At the same Fricker keeps repeating, in various phrases, how hard it is to become aware of those residual prejudices. This suggests that in her view, the problem (or the solution) of practices of prejudices turns essentially on: knowing or not knowing, being aware or not aware, being conscious or unconscious (or similar epistemic-mentalist distinctions). But it ignores many other conditions that might also explain, in a different way, why practices of prejudice persist, not so much because people are ignorant about them (this may still be the case), but because certain social conditions require the continuity of practices of prejudice. Fricker does mention, in the introductory chapter of her book, the relevance of such notions as social power and its function to 'effect social control'. But she does not use these notions to systematically explore the notion of prejudice in a way that does not put the knowing-not knowing distinction at the centre of the analysis.

3.5. Prejudice as ignorant resistance

Fricker herself suggests an alternative way of analysing prejudice. She defines prejudice not only as a judgement (whether as a belief or residue) but as resistance as well.³⁷ This follows from her interpretation of a story Arpaly tells to illustrate how prejudiced ignorance differs from 'pure ignorance'.³⁸ The story is about Solomon 'who believes that women are not half as competent as men when it comes to abstract thinking, or at least are not inclined toward such thinking. Solomon's evidence for his belief is the fact that all the women in his community, despite his attempts to engage them in learned conversation, seem to discuss nothing but gossip, family, and manual work, that the few people in his community who are interested in abstract thinking are all men, that no one he knows of has ever doubted that women are bad abstract thinkers and that the community's small, outdated library contains abstract work written by men only'.³⁹ Arpaly concludes that Solomon's beliefs are false but 'not particularly irrational because Solomon is not exposed to striking counterevidence to it'.⁴⁰ This is what Arpaly calls 'pure ignorance'.⁴¹ If, however, in another version of the story, Solomon is confronted with counter-evidence (meeting brilliant female students) and still holds to his prejudice, Arpaly concludes we should regard him as irrational. In her interpretation, Fricker concludes that this version reveals not only Solomon's irrationality but also his 'motivated irrationality, where the motivation (presumably some sort of contempt for women) is ethically noxious' (2007:34).

37 She is not the first. Cf. Allport (2005); Moody-Adams (1994); Arpaly (2003).

38 This phrase is Arpaly's: see (2003:103).

39 Arpaly (2003:103–104).

40 Arpaly (2003:104).

41 Arpaly still seems to have in mind discursive prejudice. Solomon believes that women are bad abstract thinkers, but the story does not tell us how he behaves towards women.

Fricker's shift towards cases of resisting counterevidence opens another way of looking at prejudice. As we have seen, the epistemic approach usually focuses on analysing prejudice (as a certain type of judgement) and assumes that this provides the explanation for what prejudiced people *do*: performing practices of discrimination. But here we have a type of doing that is *not* an act of discrimination: practices of resisting counterevidence. It raises questions about the underlying motivation. Does prejudice *in the sense of* a certain type of judgement also explain practices of resistance? Or does it indicate a different, not yet explored, aspect of the veil of prejudice?

Both Fricker and Arpaly seem to cling to an epistemic approach. This is indicated by their choice of connecting prejudice to (ir)rationality. Arpaly introduced the Solomon case to distinguish between prejudice as an 'honest mistake' and prejudice as ethically flawed. In the first case, the prejudiced belief would not be prejudice but simply a belief, as people would have no reason to adopt a different belief. In my view, the distinction is part of an attempt to fit a different type of doing – practices of resistance (as different from practices of discrimination) – in the mould of the epistemic approach. If people choose to persist in their incorrect beliefs 'in the face of manifest counter-evidence', this must reveal not only that their belief (i.e. their prejudice) is ethically flawed but also irrational. It follows that practices of resistance need no additional explanation but instead can be explained in terms of prejudice (again, in the sense of a certain type of judgement). In other words: the attitude to resist counterevidence is not a distinct motivation but an *integral* part of the prejudice (*idem*).

The strategy of linking prejudice to rationality is problematic for various reasons. The use of the term 'counter-evidence' seems to suggest that the (ir)rationality of a stereotype can be assessed in a matter-of-fact kind of way. Solomon either meets women who are capable of abstract thinking, or he does not. If he does meet them and clings to his stereotypical judgement, he is irrational. But assessments of (ir)rationality can be a very ambiguous and complex affair. This is illustrated by how IQ-tests have been used in the United States: the outcomes of scientific research, average differences in IQ between African Americans and whites, were used to justify claims about the superiority of whites over African Americans. But the assessment of IQ methodologies and the outcomes of IQ-tests – what is a good methodology, what does a statistical outcome really mean? – are complex matters.⁴² Or consider once again the case of Solomon. His ideas may be linked to the secondary social status that women have in their society. This status may even be part of a metaphysical or religious view on the natural order in the world: their status is considered secondary not just in a social sense but even in an ontological sense. And perhaps this metaphysical or religious view is expressed in a series of books that the Holy Men keep in the temple of the Great Male Deity and which only they may read. Once he encounters, at a university in a foreign country, female students who

42 For a critical analysis of this and other examples of scientific research that has been hampered by prejudices: see Gould (1996).

exceed even his abilities in abstract thinking, is it irrational for Solomon *not* to change his beliefs? This adapted story on Solomon suggests that beliefs – in this case, Solomon's beliefs about women – may be embedded within larger webs of beliefs. Giving up on one's beliefs then requires giving up on many more beliefs. It means that Solomon has an epistemic interest not to give in that easily to counterexamples. The adapted story also shows that what is rational is not only what one person assesses on her own but depends as much on people she 'authorises', people she trusts as having the right kind of beliefs. In Solomon's case, the Holy Men fulfil the role of epistemic authorities. Is it rational for Solomon not to believe them?

Linking prejudice to rationality also results in merging assessments of rationality with moral qualifications. This is misleading. The reasons that people have for their discriminatory practices, although they are not good from a moral perspective, may be 'good' – in an instrumental sense – from other perspectives. Focussing on the epistemic conditions of prejudices – they are false, incorrect, irrational – narrows prejudices down to something that is incomprehensible: if they are irrational, then prejudiced people must be mad?

Fricker does not consider the possibility that resistance against counterevidence might be understood as a conscious, self-chosen act that is available to people's introspection. This is indeed what theorists such as Moody-Adams seem to assume. They usually focus on cases in which people have knowingly brought themselves into a state of ignorance. Culpable ignorance is, therefore, 'essentially a matter of choosing not to be informed of what we can and should know' (Moody-Adams 1994:301). But here, 'ignorance' only means that people are ignorant of certain beliefs as being incorrect (the epistemic element). They are not ignorant of having brought themselves in that state of ignorance, i.e. of having resisted the opportunity of acquiring some information (the motivational element). Whereas the epistemic element represents a state of knowledge (not knowing that x is the case), the motivational element represents not just a state but an 'act' that consists of resisting or ignoring certain evidence from one's motivational inference processes. This includes both actions (ignoring information that might tell him that x is the case) and omissions to act, i.e. to acquire certain knowledge (not looking for information that might tell him that x is the case). My suggestion is that even resistance against counterevidence may manifest itself as a form of ignorance. If people are ignorant, not only of their irrational judgements but also of their resistance to counterevidence, this is what might be called 'ignorant resistance'. Ignorant resistance implies that people are ignorant in two ways: ignorant of having prejudices (as incorrect negative stereotypical judgements) and ignorant of resisting counterevidence. If this notion makes sense, it may explain why people, in the presence of counterevidence, display actions that resist counterevidence and, nevertheless, at an introspective level, are committed to unprejudiced views.

3.6. Ignorant resistance as a social practice of censorship

In the previous two sections, I have suggested several problems with the mentalist approach. The thread running through them is that, while prejudices are understood to be *social* in the sense that they involve a negative stereotypical judgement (in whatever sense) *about people* from certain social groups, they are not analysed as *social* in the sense that prejudices are *held by people* as belonging to certain (other) social groups. In part, this is inherent in the mentalist approach, I think, which invites analysing prejudices as judgements by focussing on (imagined) individual cases.⁴³ But it has the disadvantage of neglecting the question of how prejudices (as judgements) are shaped and reproduced in the interactions of discriminating people among themselves and in their interactions with discriminated people. In this section, I want to pick up on this question by turning to Bourdieu's concept of habitus, which is appealing for its promise to conceptualise prejudice as both invisible to prejudiced people and as a social phenomenon. A few others have also recognised the importance of Bourdieu's concept for understanding prejudice.⁴⁴ I believe my approach sufficiently differs from theirs as I will use his concept of habitus to connect prejudice (ignorant resistance) to Bourdieu's views on censorship.

Bourdieu's notion of habitus dates back to the period of Bourdieu's research into marriage in Algerian-Berber societies.⁴⁵ Structuralists before him had relied on their informants' accounts and concluded that the predominant type of marriage was marriage to a parallel cousin. By doing statistical analysis, Bourdieu discovered that the marriage patterns involved in the informants' practice were different from and did not correspond to the informants' accounts. The supposedly predominant type of marriage appeared to account for only 3 to 4% of all cases. It made him realise that there is a distinction between what people do, their practices, and what they say they do, the ideology, and that ideology need not correspond to the actual practice. Although informants are capable of 'quasi theoretical reflection', Bourdieu believes these accounts should be distrusted: 'this learned ignorance can only give rise to the misleading discourse of a speaker himself misled, ignorant both of the objective truth about his practical mastery (which is that it is ignorant of its own truth) and of the true principle of the knowledge his practical mastery contains' (1977:19). The dissonance between ideology, the result of the quasi-theoretical reflection, and practice, is what Bourdieu calls the problem of officialisation: 'officialization is only one aspect of the objectifying process through which the group teaches itself and conceals from itself its own truth, inscribing in objectivity its representations of what it is and thus binding itself by this public declaration' (1977:21).

The problem of officialisation stimulated Bourdieu to develop the concept of

43 See, of course, the Solomon case, as introduced by Arpaly (2003) and discussed by Fricker (2007). Also, see, e.g. the case of *Juliet, the implicit racist* (Schwitzgebel 2010).

44 Cf. I.M. Young (1990b); Trepagnier (2010); Shotwell (2011); Cope (2008).

45 On habit, habitus and Bourdieu's reasons for choosing this term: see chapter 6, section 3.

habitus. It is a system of structures that people have internalised in their attitudes, not only in how they act and react towards people from other social groups but also in how they perceive and understand their social world. Habitus arises as people's response to the objective conditions of existence (economic, social, et cetera) that build up the social world in which people grow up. Past experiences condition people to respond in particular ways to these conditions. Whenever similar conditions occur, people will therefore perform the same reactions. When conditions change over time but do not compel them to adapt their habituated responses, people may persist in the same reactions. Although people's habitus, as a structural, internalised response to objective conditions, lends objective social meaning to their practices, this does not mean that people perform these practices with the intention of realising this objective social meaning. As such, the habitus is 'the source of [a] series of moves which are objectively organized as strategies without being the product of a genuine strategic intention' (1977:73).

For her analysis of epistemic injustice, Fricker has considered using the notion of 'social imaginary', for which she refers to Gatens (1996) and to the work of Cornelius Castoriadis, but in the end, she has chosen for 'the less heavily theoretical notion of the social imagination' (2007:38, note 9). It may explain why it is difficult to understand from her text how she relates prejudices to the social imagination. She sometimes tends to reify 'social imagination' as something that stands somewhere outside the individual. She might have found Bourdieu's notion of habitus useful: although Bourdieu does not analyse habitus in terms of 'imagination', it does express some sort of 'collective knowledge', or what we might call, in Bourdieu's case an appropriate metaphor, a 'body of knowledge' that people share.⁴⁶ For Bourdieu, the habitus is not something outside people; it is embodied in what they do, what they say (or better: *how* they say what they say), how they perceive. It seems reasonable to assume that habitus also includes prejudices. But we should be careful about understanding prejudices along the lines of Fricker's epistemic approach. Various theorists have observed that Bourdieu's habitus is congenial to Gadamer's *sensus communis*.⁴⁷ Notions like habitus and *sensus communis* share the idea that much (perhaps *most*) of the knowledge that people share when they live in the same community is not available as explicit, conscious, propositional

46 Using Bourdieu's habitus to support Fricker's analysis on epistemic injustice has previously been suggested by Medina (2013: 269–270), but in the end, he preferred using the work of Cornelius Castoriadis.

47 See, e.g. Holton, who remarks: 'In some ways, Bourdieu's use of the term common sense to refer to the sense of reality of specific communities is similar to Gadamer's (itself borrowing from Vico). It seems, however, to be almost a relationship of mirrored opposition: while Gadamer emphasizes the enabling aspects of *sensus communis*, Bourdieu emphasizes its limits' (1997:47). Also, see King (2000), who seems to suggest that habitus should not be confused with the notion of life-world (*Lebenswelt*), as for example, Habermas uses it, as this notion is part of a dualistic view on social structures/ knowledge, a view that King rejects. Bourdieu's habitus also reminds of Gadamer's concept of prejudice (*Vorurteil*), not in its current sense as a negative judgement, but in its original sense as 'pre-judgement'. Gadamer (1990:276–281; or: second part, II, chapter 1, section a, subsection β).

knowledge.⁴⁸ I take it that habitus, for that reason, implies some form of ignorance. Bourdieu denies that people can make their habitus explicit. If people's habitus also includes their prejudices, we should therefore not understand them as conscious (or even unconscious) judgements. Of course, this raises the question of how we should understand this type of knowledge. In the introduction to this chapter, I have already indicated that I will sidestep this question, in this chapter at least. For now, it suffices to assume that habitus, apart from all other kinds of 'beliefs', including stereotypical 'beliefs' about other people, may also include prejudice (as ignorance). Habitus functions as the motivating force behind people's social practices. If habitus includes prejudices, then it may therefore also explain social practices of discrimination.

Interestingly Bourdieu has explored the relations between habitus and practices of censorship. Depending on circumstances, people may come to experience considerable advantages if they, as a social group, perform practices that exclude, marginalise, or disempower other people by treating them as belonging to another social group. Over time these practices take on a certain, stable pattern that corresponds with – in terms of internal motivations and judgements – a prejudiced habitus. People from social groups, or perhaps certain subgroups, who perform practices of discrimination, may have an interest in keeping the habitual negative judgements that sustain the social structure that is favourable to them: 'the habitus tends to ensure its own constancy and its defence against change through the selection it makes within new information by rejecting information capable of calling into question its accumulated information, if exposed to it accidentally or by force, and especially by avoiding exposure to such information'.⁴⁹ Their habitus will therefore motivate them to preserve as much as possible the way that all people in society perceive the societal order. In the discussion on Fricker's views, the notion of 'counter-evidence' was already introduced. I will use here in a broader sense, as including not only facts but also counterviews or critical practices. Counterevidence may function to question certain social structures that are favourable to discriminating people. If this is the case, these people may therefore adopt an attitude of resisting this counterevidence. Discriminating people may choose between at least two different strategies of resisting the counterevidence. They may choose to simply *contradict* any counterevidence that discriminated people put forward. They may also try, which may prove to be more effective, to prevent counterevidence from even being put forward in the first place. While the first strategy would consist, probably, in active participation in public debate, where the counterevidence and the counter-counterevidence are manifest for all, the second strategy is less manifest, more secretive because it would consist in 'excluding certain agents from communication by excluding them from the groups which speak or the places which allow one to speak with authority'.⁵⁰

48 In modern philosophy, a great variety of proposals have been made to make sense of this type of knowledge (which I here refer to only in what type of knowledge it is *not*).

49 Bourdieu [1980] (1990):60–61.

50 Pierre Bourdieu (1991:138).

Bourdieu understands this social practice as ‘censorship’. At first, it is manifest ‘in the form of explicit prohibitions, imposed and sanctioned by an institutionalized authority’.⁵¹ Only later, it gradually transforms into another type of censorship that each agent internalises in such a way that he experiences it as something natural. In its most effective form, this agent ‘does not even have to be his own censor because he is, in a way, censored once and for all, through the forms of perception and expression that he has internalized and which impose their form on all his expressions’.⁵² As a result, prejudice itself will be obfuscated within public discourse itself.

Bourdieu and Fricker share a distrust in what people believe or say they believe. They both draw attention to certain internalised, hidden motivational forces, either as residual prejudice or as habitus. But I think that Bourdieu’s habitus can provide a richer notion of prejudice. While Fricker’s residual prejudice may explain the distortion of people’s perception, resulting in epistemic injustice, Bourdieu’s habitus is able to explain not only this type of distortion but also why practices of prejudices persist, namely as the result of practices of censorship. Even more so, as Bourdieu will claim, habitus also explains how it is possible that people, both the discriminating and the discriminated, are ignorant about practices of censorship: as the result of a social process of internalisation, they forget about practices of censorship *as censorship*.

3.7. Social practices of silencing

Implicit in Fricker’s and Bourdieu’s analyses is a distinction between two different kinds of practices: discrimination and silencing. They are related to the two cognitive-psychological elements of prejudice: a negative stereotypical judgement *and* an attitude of resistance against counterevidence. Negative stereotypical judgements motivate people to discriminate, which disadvantages people of certain social groups in an objective way (be it materially or symbolically). Although the attitude of resistance cannot occur without prejudice as judgement, it does motivate people to embark on practices that are distinct from discriminatory practices: they form a different class of practices. Fricker qualifies part of this class as practices of testimonial injustice. Bourdieu calls them censorship. Here I call them ‘practices of silencing’ as they are intended to suppress the occurrence of any view or evidence endangering the institutions that enable those practices of discrimination to continue. As the analyses of Fricker and Bourdieu show, people may be silenced in many ways: from discrediting people’s credibility to excluding them from the groups which are allowed to speak. Practices of prejudice comprise not only practices of discrimination, originating in prejudice as a judgement (either as an implicit association or as part of people’s habitus), but also practices of silencing, originating in an attitude of resistance. Practices of silencing function to keep

51 Ibid.

52 Ibid.

intact those social structures that are favourable to discriminating people (the dominant, as Bourdieu would say). It also means that practices of discrimination are kept intact.

Fricker's concept of testimonial injustice and Bourdieu's concept of censorship as habitus nevertheless do not yet adequately capture the meaning of silencing practices. Fricker's concept has a focus on assessments of credibility. It also lacks, as I already argued, a view on the reproduction of prejudices. Bourdieu's concept of habitus helps to explain not only how individuals' behaviours (including perception) originate in social practice but also why they are reproduced. But in applying habitus to this phenomenon, Bourdieu does not fulfil the promise of his own notion. To explain this, I will first argue that Bourdieu's analysis fails to reflect on the concept of sanctions which is, I believe, inevitably connected to censorship. I will then argue that his notion of 'internalised censorship' can account only for prejudice as judgement, not for prejudice as resistance. Next, I will argue for a revision of Bourdieu's censorship by arguing for the notion of naturalised sanctions. The result is a special variety of censorship: silencing, which operates not through explicit prohibitions but through sanctioning practices that people are ignorant of.

My suggestion is that we understand the concept of censorship as consisting of two layers: censorship itself and sanctions. Censorship itself implies norms that tell people what to criticise or not in society ('Do not tell others about what our company is doing'), whereas sanctions imply norms of how to deal with violations of these norms ('And if you do, you will get fired'). One function of sanctions is to instruct people on what a correct practice is, a function which is important in educational contexts, for example, when parents reprehend their children for behaving badly. Another function of sanctions, a distinctive one for censorship, is to inculcate people what a correct practice is. They may have learned and understood the contents of the norms due to instructive sanctions, but this does not imply that all people who have incorporated these norms judge them to be right. They may disagree and lose commitment to them. Censorship can only be effective if sanctions are implemented consistently. When the dominant in their relationship with the dominated do not put sanctions to use, or when sanctions are non-existent, the non-dominant will get a chance of voicing their criticisms of discriminatory practices and of influencing the views of the dominant.⁵³

53 The dominant-dominated distinction is part of Bourdieu's vocabulary. I understand these terms and their contrast to mean that the dominant are those people who are, compared to dominated people, generally in more favourable conditions (socially, financially, culturally, et cetera) due to (gender, financial, historical, ethnical, et cetera) characteristics that define these people as belonging to (various, intersecting) social groups. This probably begs for more explanation. For my explanation here, it suffices that the dominant are identified with those who perform practices of prejudice *and* are generally and more often than the dominated in positions of (cultural, political, financial, media, et cetera) power and influence, while the dominated are identified with those who suffer from those practices of prejudice.

Bourdieu does not distinguish in this way between censorship and sanctions. Applying this distinction to his analysis helps to reveal the weakness in his concept of censorship. Explicit prohibitions can be understood as expressing norms of censorship combined with practices of threats and, if necessary, the imposition of sanctions. Bourdieu suggests that in due time, assuming that sanctions are stable and consistent, people, both the dominant and the dominated, experience the norms of censorship as being natural. This may seem the most effective form of censorship as it makes sanctions become obsolete. But I believe this notion of internalised censorship is self-defeating. Sanctions form the crucial link between norms of censorship and the preservation of the societal relations of power that are favourable to the dominant. If we assume that as soon as people start to experience censorship as natural, sanctions are no longer needed, the dominant will start to forget that the norms of censorship had a purpose in preserving the existing relations of power. As for the dominated people: even if now they are committed to the internalised norms of censorship, this commitment will nevertheless, in the absence of sanctions, gradually lose hold on them because they still feel the need to question the practices of discrimination they keep suffering from. Internalised censorship, as Bourdieu understands it, therefore creates the very opportunity of doing what should have been censored. This notion should include both norms of censorship and sanctions.

The problem with Bourdieu's reconstruction of the process of habituating censorship is that he still starts from a classic notion of censorship as 'official legal suppression'⁵⁴ which is directed at suppressing explicit counterviews. Although his analysis of habituating censorship may be right for some societies and for some practices, it cannot account for the process of hidden prejudices that get institutionalised. Fricker's notion of resistance is more appropriate. I briefly return to the notion of 'views'. Previously I redefined the term counterevidence as also referring to (counter-)views, as I wanted to avoid that prejudiced judgements were understood only in terms of truth-values. Views may be explicit in the form of ideologies or religious views. We already concluded that Bourdieu's habitus helps us to show that people may be ignorant of their prejudices but also of other kinds of views they have. Views may therefore also refer to 'ideologies' or, as Fricker would call it, 'social imaginaries' that are implicit and hidden.⁵⁵ Analogously counter views can also be distinguished into explicit and implicit. I believe the process of habituating censorship should be understood as an interaction between implicit views and implicit counterviews.⁵⁶ The process starts off

54 Cf. Post (1998a).

55 The practices performed out of habitus have a meaning that is hidden in the sense that people have very little reflexive understanding of their practices, but that is also not hidden in the sense that people are able to understand and respond, in a pre-reflexive way, to each other's actions. In the following paragraphs, I will talk about these responses. To avoid confusion on the term 'hidden' I will therefore use only the term 'implicit'.

56 For a historical illustration of such interactions: see Scott (1990).

with occasional practices of discrimination gradually forming into habitus, as objective conditions are favourable to forming this habitus, while this process elicits practices of resistance (implicit counterviews) which, in turn, elicit a counter-practice of silencing. To give a general idea of how this process works, I will first home in, with the use of Bourdieu's own notion of habitus, on the experience of people who are subjected to discrimination and how they internalise both censorship and sanctions; and will also focus on the experience of the dominant who develop a strategy of implicit censorship directed at implicit counterviews.

The experience of those who are subjected to persisting practices of discrimination should be reconstructed as the internalisation of censorship and sanctions and as a source of (implicit) counterviews. To illustrate this, I will briefly turn to the experiences of a fictional character. Mark Twain's story 'Adventures of Huckleberry Finn' has captured the attention of philosophers because of Huckleberry Finn.⁵⁷ In terms of how prejudice operates, I believe it is also interesting to consider how Jim, a runaway slave and Huckleberry's companion during his adventures, is affected by the prejudices of their society. At one point in the story, Huckleberry finds that Jim has run away from Miss Watson, who owns him as a slave. Jim explains: 'Ole Missus — dat's Miss Watson — she pecks on me all de time, en treats me pooty rough, but she awluz said she wouldn' sell me down to Orleans. But I noticed dey wuz a nigger trader roun' de place considerable lately, en I begin to git oneasy. Well, one night I creeps to de do'pooty late, en de do' warn't quite shet, en I hear old missus tell de widder she gwyne to sell me down to Orleans, but she didn' want to, but she could git eight hund'd dollars for me, en it 'uz sich a big stack o'money she couldn' resis'. De widder she try to git her to say she wouldn' do it, but I never waited to hear de res'.⁵⁸ He then runs away. Later in the story, his motives become even clearer: '[Jim] was saying how the first thing he would do when got to a free State he would go to saving up money and never spend a single cent, and when he got enough he would buy his wife, which was owned on a farm close to where Miss Watson lived; and then they would both work to buy the two children, and if their master wouldn't sell them, they' get an Ab'litionist to go and steal them.'⁵⁹ The experiences of Jim, which Mark Twain partly tries to capture here from Jim's own perspective, are exemplary, we might say, of those African Americans who have internalised their discrimination as natural. African American people became part of a society that acquired the habitus to categorise them as slaves and treat them in a discriminatory way. In the same way, their own responses, obeying and giving in to these practices of discrimination, became habitual to the point that black people had internalised the images and views of the dominant and perceived practices of discrimination as natural, as part of a social world to which they, as a social group, could offer no alternative view.

57 See e.g. Bennett, (1974); Arpaly (2003); Fricker (2007).

58 Mark Twain (1982:666).

59 Mark Twain (1982:711).

The experience of being discriminated does not stop at the internalisation of discriminatory practices but also includes feelings of discontent. If practices of discrimination continue to exist, the dominated continue to also experience these disadvantages of these practices to themselves. This experience, which translates into feelings of discontent or, as Jim feels it, uneasiness, is a source for developing counterviews. I hesitate here to choose words as 'discontent' or 'uneasiness', as they articulate what the dominated, in any historical or social setting, should themselves articulate. The articulation of this experience is already a step toward formulating an explicit counterview. But even before this, the dominated can express a counterview that is implicit in what they do. Mark Twain's story shows how Jim once accepted his situation as a slave. By running away, Jim implicitly criticises being held as a slave. Although such acts do not necessarily reflect a counterview in which a dominated individual self-reflectively generalises his own experience, it nevertheless expresses a counterview. The habitus the dominated develop in internalising censorship turns out to be a complex one. The dominated experience the implicit views of the dominant as natural and simultaneously experience the disadvantages of the dominant social structure, which is a source of counterviews, displayed implicitly in practices or explicitly in articulated counterviews. This analysis, however concise, indicates that the habitus of the dominated is more complex than Bourdieu's own analysis of internalised censorship can account for.

The experience of uneasiness, or discontent, will remain a source of actions of resistance, whether openly or secretly. Even the dominant, although they do not share the experience of being dominated, will at least be sensitive to the potential of resistance, for which reasons they sense the need of upholding implicit sanctions. This will elicit from the dominated resistant counteractions. This implies for the dominant the need to continue sanctions which the dominated, in their turn, internalise. Here we have a dialectical process in which the habitus of the dominant differs from the habitus of the dominated. Of course, there is a tension between, on the one hand, experiencing practices of discrimination and sanctions as natural and, on the other hand, experiencing those practices as burdensome and uneasy. But this tension is a real one, not a conceptual one: people are able to experience certain practices as natural and still feel unhappy about them. The implicit practices of resistance that dominant social groups perform to respond to this quality as practices of silencing.

The veil of prejudice – i.e. dissonant practices of prejudice – appears to be a complex phenomenon. The hidden prejudice, embodied in people as either implicit bias or habitus, may be understood as a motivating force for performing two different kinds of practices: not only *practices of discrimination* but also – and perhaps more fundamentally – *practices of silencing* that preserve the continuing existences of practices of discrimination. Practices of silencing indicate an underlying attitude of resistance against 'counter-evidence', which I understand here in a broad sense, referring to facts,

claims, practices and evaluative views that in whatever imperfect way tend to reveal certain social practices as practices of discrimination.

3.8. Conclusion

In this chapter, I have presented prejudices as a case of (causing) agential self-ignorance: the veil of prejudice. First, I explored how we may understand that people are ignorant of their hidden prejudices. For this, I chose the notion of implicit bias, as developed in recent decades by social psychologists. I then turned to the question of how hidden prejudices may be understood as a social phenomenon. I discussed two different approaches: a mentalist-epistemic one (Fricker) and a pragmatist one (Bourdieu). While the notion of (hidden) prejudice is usually invoked to explain discriminatory practices, both approaches try to move beyond it and understand hidden prejudice as accounting also for practices of silencing (what Fricker calls ‘epistemic injustice’ and Bourdieu calls ‘censorship’). In a critical discussion of both approaches, I made some suggestions for an improved notion of the veil of prejudice as a social phenomenon.

I have treated both approaches as if they are largely compatible, but I have not attempted to integrate them as they start from basically different starting points. The epistemic-mentalist approach defines prejudices in internal terms, i.e. in terms of unconscious states, and as the product of psychological, equally unconscious processes. Bourdieu has explicitly criticised the mentalist approach (in sociology).⁶⁰ Instead, his pragmatist approach understands prejudices in *external* terms, i.e. of what people *do* and as the product of social processes. Deciding between a mentalist or pragmatist approach will be deferred to chapters 5 and 6. For now, I will not yet take any side.

My aim with this chapter was to design a specifically social-political version of agential self-ignorance because I want to put the liberal-mentalist notion of introspective self-knowledge to the test, but preferably in a context that is relevant to my research question. The veil of prejudice qualifies as such. My next move will be to take up the challenge that was left behind in the second chapter: is the liberal-mentalist notion of responsibility able to deal with agential self-ignorance? To explore this question, in the next chapter, I choose the social-psychological notion of implicit bias, as this notion relies on a mentalist framework, as does the liberal-mentalist notion of responsibility.

⁶⁰ Bourdieu developed his notion of habitus as an alternative to (what I call) the mentalist vocabulary. In an interview, he stated explicitly that consciousness and unconscious are ‘the very alternatives that the notion of habitus is meant to exclude’ (Bourdieu 1990:10). In chapter 6 (section 3), I will explore his reasons for choosing this strategy.

4

Chapter 4

**The mentalist framework:
the impossibility of responsible agency**

4.1. Introduction

Raskolnikov, the well-known protagonist in Dostoyevsky's novel *Crime and Punishment*, murdered a pawnbroker, an old woman, in early July of 1865. The name of the fictional woman is Alyona Ivanovna. Raskolnikov believed that society consists of ordinary people for whom the moral law obtains and extraordinary people that may transgress the boundaries of moral laws if their acts benefit society. He was convinced that he belonged to this second class and that his act, for that reason, was justified.¹ But we have no need for fictional examples. In early January 2012, a woman was killed in the quarter Jarov, the third district of Prague.² She had been kicked, punched in the head and then was stabbed with a knife. It was an unremarkable murder. She had been homeless and a Romani woman. In an attempt to gather more information from the Prague police, the European Roma Rights Centre was refused the exact date of the crime.³ Neither does the available information reveal the name of the Romani woman. She remains anonymous. All we know is that she was murdered by three young men, neo-Nazis, who were known to have attacked homeless people and locals before. A Czech legal organisation that monitors hate violence, *In Iustitia*, was asked for a comment on the crime and explained that 'brutal attacks on the homeless were becoming more frequent and the neo-Nazis mindset is 'that they are performing acts beneficial to society as a whole' (Fekete 2012:16).⁴

This incident illustrates responsibility for prejudices in several ways. Murder belongs to the basic categories of moral and penal wrong. So, of course, it might be analysed as the responsibility the young neo-Nazis have for their crime. From a juridical perspective, the issue of their prejudices and how they fit in a broader societal context may not be that interesting. What is relevant in a juridical sense is that a murder has been committed and to what extent it can be shown that it was committed by these young men. From the moral perspective that I have chosen for my argument, it is relevant whether they have prejudices, and even more so, hidden prejudices. And indeed, these men might be said to have several prejudices: against women, against homeless people and against Romani people. The incident allows for homing in on yet another class of actors: the policemen or if we take a broader perspective, Czech or even European officials. It would be more

1 'Kill her, take her money, and with its help devote yourself to the service of humanity and the good of all. And, when you come to think of it, what does the life of a sickly, wicked old hag amount to when weighed in the scales of the general good of mankind? It amounts to no more than the life of a louse or a black beetle, if that, for the old hag is really harmful.' The quote is taken from Bergelson (1996:924).

2 See <http://www.romea.cz/en/news/czech/romani-woman-murdered-in-prague-locals-allege-perpetrators-are-nazis>.

3 European Roma Rights Centre et al. (2012).

4 Dostoyevsky associated this type of reasoning to which Raskolnikov was also committed with utilitarianism (Offord 1983). He was an early critic of utilitarianism and its naturalistic outlook, expressed in its commitment to what Hacking (1983) calls statistical determinism, which provides another variety of the one discussed in this chapter. More on *In Iustitia*, see <http://en.in-ius.cz/about-us>.

interesting to analyse their responses as displaying hidden prejudices against women, the homeless and Romani.⁵ Do they have prejudices and, if so, are these perhaps examples of hidden prejudices?

In the previous two chapters, I have introduced the liberal-mentalist notion of responsibility (chapter 2) and the veil of hidden prejudice (chapter 3). My aim in this chapter is to pick up on its psychological version, *implicit bias*, and ask whether people are responsible – according to the liberal-mentalist notion of responsibility – for their implicit biases. What is different about the psychological approach, as compared to the Bourdieusian-pragmatic approach, is that the *first* takes hidden prejudices to be *essentially* a *natural* phenomenon while the *second* takes them to be *essentially* a *social* phenomenon.⁶ Its interpretation of implicit biases as a natural phenomenon is a *neurological* one: implicit biases are rooted in the neurological hardware of people's brains. The psychological approach is itself part of a broader theoretical framework that includes a deterministic view on behaviour. My claim in this chapter is that if we accept the neurological interpretation of the conscious/unconscious distinction and thereby accept its underlying commitment to a deterministic world view, this provides the arguments for a defeat of the liberal-mentalist notion of responsibility.

Linking the debate on responsibility for implicit bias with the debate on determinism means treading new ground in more than one respect. First, the current debate on responsibility for implicit bias itself is concerned with introspection as self-knowledge, but not with determinism. On the other hand, the standard debate on determinism, often also known as a debate on 'moral responsibility', is about the problem of responsibility as control over one's actions versus determinism but is not concerned with responsibility as either self-knowledge or consciousness.⁷ Finally, debates on self-knowledge and those on consciousness are also not commonly linked.⁸

In the next section, I will discuss three claims on responsibility for implicit biases and explain why I choose, for the argument in this chapter, to stay away from the 'local' problem of *responsibility for implicit bias* and instead focus on the 'global' problem of *responsibility in a deterministic world* (section 2). The metaphor of manipulation is frequently used in this debate to illustrate or expose intuitions on determinism.

5 In the case of the homeless Romani woman, it is not clear whether any of the stereotypes (woman, being stereotype, Romani-ethnicity) had a distinctive force in eliciting prejudiced responses by the Czech police.

6 The distinction into either 'social' or 'natural' may seem a theoretically rather too rigid dichotomy. In this and the next chapters, it will appear that this distinction has its nuances. For now, I will simply take this dichotomy as a helpful starting point for my discussions.

7 For the absence of ontological discussions in the debate on moral responsibility: see, e.g. King and Carruthers (2012:200) who remark that the 'question about the relationship between theories of responsibility [i.e. 'moral responsibility' as it is discussed in debates on free will], on the one hand, and a commitment to conscious attitudes on the other [...] 'has rarely been raised previously'. Also see Caruso (2012:130-131, note 1).

8 See Smithies and Stoljar, who observe that 'the relationship between consciousness and self-knowledge has been curiously neglected and remains poorly understood' (2012:3).

Homing in on this metaphor will reveal certain assumptions that are best understood in terms of the conscious/unconscious distinction. I will explain how this metaphor provides an argument against *reflective volitional autonomy* (see chapter 2, section 3): the Manipulation Argument (section 3). Next, a case will be discussed as posing a challenge to the Manipulation Argument (section 4). This will lead to developing an argument against *reflective epistemic autonomy* (see chapter 2, section 5): the Manipulation Argument Revisited (section 4). It will conclude my attempt to show the vulnerabilities of a liberal-mentalist notion of responsibilities.

4.2. The determinism thesis

In this section, I will briefly discuss three claims about responsibility for implicit biases:

1. People are not responsible for their discriminatory behaviour as it is motivated by implicit biases that are in principle unconscious and therefore unknown to them.⁹
2. People's discriminatory behaviour is not open to *moral* change as these implicit biases are rooted in the neurological hardwire of their brains.
3. People cannot ever be responsible: people are not free agents (they are determined) because the neurological hardwire of their brains that regulates all their behaviour is unconscious and therefore unknown to them. And responsibility for any action is therefore impossible.

The first claim provides a serious challenge, but not one that cannot be overcome if these implicit biases can be changed. Therefore, the second claim poses a more serious threat. What use would it be to aspire for egalitarian justice if people cannot change their discriminating behaviour in a morally relevant sense? But the third claim poses the most serious challenge. Whereas the first two claims are potentially undermining for processes of realising justice, the third claim seems to destroy the very project of realising justice because it denies people their free agency and, therefore, their responsibility. If we take the third claim seriously, we should deny responsibility not only to the Prague police for any hidden prejudice they might have had in conducting the research on the murder against the Romani woman but even to the young neo-Nazis that committed that murder.

I start with the first claim. Research in social psychology indicates that people have no introspective access to their implicit biases. While research on implicit bias is rapidly growing, less attention has yet been given to exploring the moral implications of implicit

9 What supports this view is 'The 'Simple Argument', as Zheng calls it (2016, section 3). As representatives of this view, he mentions Saul (2013) and Blum (2004).

biases.¹⁰ The few analyses that have are usually structured by the liberal-mentalist notion of responsibility in taking reflective control as the defining feature of responsibility.¹¹ The main conclusion to be drawn from these analyses is that ‘individuals are not responsible, and therefore not blameworthy, for their implicit biases, and that this is a function of the nature of implicit bias as *implicit*: below the radar of conscious reflection, out of the control of the deliberating agent, and not rationally revisable in the way many of our reflective beliefs are’ (Holroyd 2012:274).¹² The realisation that people might not be responsible for their hidden prejudices has troubled at least some of these authors:

‘It has led me to have the following nightmare: After testifying for the plaintiff in a case of egregious and demonstrable discrimination, a cognitive social psychologist faces the cross-examining attorney. The hostile attorney, who looms taller than Goliath, says, “Tell us, Professor, do people intend to discriminate?” The cognitive social psychologist hedges about not having any hard data with regard to discrimination, being an expert mainly in stereotyping. When pressed, the psychologist admits that stereotypic cognitions are presumed to underlie discriminatory behavior. Pressed still further, the psychologist reluctantly mumbles that, indeed, a common interpretation of the cognitive approach is that people do not stereotype intentionally, whereupon the cross-examining attorney says in a tone of triumph, “No further questions, Your Honor.” The plaintiff is led shaking from the courtroom. . . .’ (Fiske 1989:265).¹³

If people are not even aware of what they were doing when they were performing discriminatory behaviour, it seems that holding them responsible for it makes no sense.

The research results on implicit biases hold yet another, more pressing threat, which brings us to the second claim. They show that implicit biases are correlated to certain neural activities of people’s brains.¹⁴ Researchers on implicit bias seem to suggest that implicit biases are not just *indicated* by those neural activities, but somehow *are* those neural activities ‘as if

10 See, e.g. Fiske (1989 and 2004); Wellman (2007); Kelly and Roedder (2008); Kennett and Fine (2009); Machery, Faucher and Kelly (2010); Holroyd (2012); Levy (2012; unpublished manuscript); Madva (2012; unpublished manuscript); Brownstein (2013); also see Saul and Brownstein (2016).

11 The participants in the debate on responsibility for implicit bias all seem to be committed exclusively to the liberal-mentalist notion of responsibility. For some exceptions: see this chapter, note 22.

12 Pereboom summarises what seems to be the dominant view among compatibilists: ‘Many compatibilists would agree that when an action comes about as a result of covert manipulation (of the right sort), the agent will not be morally responsible’ (2001:112). Holroyd herself seems to be one of very few to claim that people are still responsible for their hidden prejudices. Also, see Brownstein (2015) and, to some extent, Washington and Kelly (2016).

13 Also, see Bargh, who asks a similar question (but seems less concerned): ‘how could anyone be held responsible, legally or otherwise, for discriminatory or prejudicial behavior when psychological science had shown such effects to occur unintentionally?’ (1999:363).

14 See, e.g. Fiske (2000); Stanley, Phelps and Banaji (2008).

stereotypes and bias are hardwired in the human brain' (2007:51), as Wellman observes in a critical analysis of the theoretical framework of research on implicit bias:

'Thus, there is no space for intervention or change, except for undefined, Orwellian notions like "mental correction" or "careful process re-engineering." [...] And if biases are inevitable, part of normal cognitive functioning, it is not possible to legislate or litigate an end to discrimination. Like implicit bias, racism is inescapable.' (Wellman 2007:51).

In the neural hardwire interpretation, it is not clear how people may be changed. And therefore, Wellman concludes:

'cognitive neuro-science considers encoded bias as fixed, automatic, inevitable, it has no theory of social change' (2007:65).

In response to this criticism, Machery, Faucher, and Kelly have argued that the neural mechanism on which implicit biases rest does not stand in the way of changing them:

'implicit biases are amenable to such innocuous methods as having participants focus on counter stereotypic imagery [...] or images of admired Black celebrities [...], or having them interact with a Black person who is in a superior or prestigious role [...]' (2010:242).

But their conclusion may be premature. The suggested methods to undoing implicit biases may perhaps bring about social change but do not necessarily agree with the requirements for *moral* change. Moral change, as the liberal model of responsibility suggests, is established because people engage in deliberative, public dialogue and are persuaded by reasons. Does exposing people to counter-stereotypic imagery or images provide an answer to the liberal model of moral change, or perhaps any other model? Will these people be told that they are going to be exposed to it? Do they need to be persuaded? Will they have a right to refuse? And if they are subjected to debiasing methods without being told, are they not being *manipulated*?¹⁵

The most serious problem, however, that research on implicit biases poses to the liberal theory of responsibility is to be found at a deeper level of analysis. The third claim adopts a naturalistic approach to the problem of responsibility for actions originating in implicit bias and generalises it as a problem of responsibility for all human actions.

15 In addition to Wellman (2007), also see Thomas and Brunsmas (2014) for an interesting critical discussion of how the research on implicit bias brings about a shift in academic discourse from viewing racism as a social or cultural problem to racism as a medical problem. This has consequences too for views on how racism can or needs to be 'cured'. Underlying the 'pathologisation of racism', as the authors call it, lies the same problem of ignoring the issue of what counts as *moral* change.

Naturalism assumes that whatever causal succession takes place between one fact and another fact, between one event and another event, it is bound by mechanical causation.¹⁶ Let us call this the Mechanical Causation Thesis. Part of the naturalist ambition is to ‘naturalise’ human actions, i.e. provide a mechanical explanation of them that ‘takes priority over, indeed renders false, any explanation in terms of desires, beliefs, intentions’ (Dennett 1973:160).¹⁷ Efforts to naturalise responsibility have been the subject of a specifically distinct debate: the determinist debate on ‘moral responsibility’ or ‘free will’. It starts from a more specific, more rigid variety of the Mechanical Causation Thesis: the Determinism Thesis, ‘which claims that the past facts, together with the laws of nature, entail all future facts’ (Campbell, O’Rourke and Shier 2004:2).¹⁸ Determinism, we might therefore say, is the control of no one in particular: it is the tyranny without tyrant of all matter over itself. Naturalists engaged in the debate on naturalising responsibility are mostly ‘determinists’, recognising the Determinism Thesis. They are divided into those who believe that responsibility and determinism are compatible, the ‘soft determinists’ or ‘compatibilists’, and those who deny this, the ‘hard determinists’.¹⁹

The current debate on implicit bias does not include a discussion of the third claim.²⁰

-
- 16 One way of understanding the disagreement between naturalists and non-naturalist is to take them not to disagree on the question of whether some sort of correlation exists between body and mind, between physical states and mental states, between neural processes and thought. Instead, their disagreement is on how to understand this correlation, i.e., what has priority in the order of explanation? In explaining the mind naturalists, claim priority for natural facts over the facts of the mind.
- 17 Dennett believes mechanism and responsibility can be reconciled. Even if we sidestep that Dennett discusses freedom in terms of mechanism instead of determinism, his position still does not fit in the determinist debate on freedom because Dennett rejects theatrical notions of consciousness and therefore must reject the notion of reflective will. Instead, he chooses the notion of freedom as rationality. I will return to this in the next chapter.
- 18 I must add that the real challenge is strictly not the determinist, but the mechanical thesis or what, for example, Scanlon calls the ‘Causal Thesis’ (1986:152). Sometimes several types of determinism are distinguished, each of which earns its own adjective. (In this chapter, note 4, I mentioned already one version: statistical determinism.) For more references on definitions of determinism: Campbell, O’Rourke and Shier (2004:2, note 2). Determinists usually pay their respects by mentioning Laplace. Laplace was probably not a determinist in the current sense because, in his view, determinism applied to matter, but not necessarily to the mind (see Hacking 1983:458).
- 19 They usually have in mind the moral-legal (i.e. basic desert) notion of responsibility. This does not make a difference to the argument in this chapter.
- 20 For the discussion in this chapter, the differences between the mechanical and determinist thesis have no relevance. Some theorists on responsibility for implicit bias seem to recognise the possible relevance of the naturalist (determinist) thesis for their debate. They, too, choose not to address it. See, e.g. Brownstein, who remarks: ‘I entirely sidestep the question of whether this emerging picture of the mind threatens free will’ (2015:3, note 10). Also, see, e.g. Washington and Kelly, who, in an article on responsibility for implicit bias, acknowledge the possibility of analysing it as part of a broader approach addressing ‘global threats to free will and moral responsibility. For instance, one such threat seems to emerge from contemporary physics, and an appreciation of the fact that there seems to be no room for genuine choice or moral responsibility if we are living in a deterministic universe’ (2016:15). They choose, however, to defer this issue by adopting a compatibilist stance. ‘Although fascinating’, they say, ‘we will adopt the assumption built into our ordinary, everyday practices of holding people responsible – typical of compatibilist approaches to free will – that we are responsible for some of our behaviors’.

None of the authors mentioned earlier, including Holroyd and even Wellman, addresses the broader problem of either the Mechanical Causation or Determinism Thesis as naturalising human agency and thereby dissolving responsibility.²¹ The reason for ignoring this problem is, probably, that they have taken the issue of naturalism already into account by tacitly choosing the compatibilist solution: accepting the Determinism Thesis but rejecting the implication that moral responsibility is impossible.²² But not addressing the third claim is a mistake, as I will explain.

Previously I suggested (see chapter 2) how the liberal notion of responsibility developed towards 'self-government' and how this required a more sophisticated moral psychology, one that was able to explain how people gain knowledge of moral principles as their own principles. It was found in the notion of reflection. In the transition from the early-modern to the modern period, as the scientific, disenchanted outlook on the world gained prominence, the mind (conscious) began to be recognised as part of nature. This allowed viewing reflection as a psychological phenomenon. At the same time, reflection was described more and more in terms of the 'conscious' and 'introspection'. But from the very start that they came into use as technical terms, they were characterised by an ambiguity: they were used both for epistemological purposes – to describe self-knowledge: the process of acquiring knowledge of mental states as one's own – and for ontological purposes (in the philosophy of mind) – to describe what kind of thing (self-)consciousness could be in a world that was disenchanted and consisted therefore only of objects. This ambiguity of the mental as both an epistemological and ontological phenomenon remains an important undercurrent in the liberal notion of responsibility.

In the debate on responsibility for implicit bias, it is mostly the epistemological dimension that comes to the fore. Implicit bias raises for people's responsibility the problem of self-knowledge: if people are responsible only for what they themselves have consciously endorsed, then how can they be responsible for any motivation they were not conscious of and cannot even become conscious of in principle? But this definition of the problem arises only for those who are committed to both the liberal notion of

21 As one of few, Durrheim (2012) points to the problem of 'the determinist and mechanist view of the prejudiced subject'. While he is critical about research on implicit bias, he does not further analyse this particular problem. Surprisingly Bhaskar, who opposes the idea of determinism as 'an immensely implausible thesis' (1975:18), does not seem to recognise the mechanistic or deterministic view underlying research on implicit bias, as he has argued, with his co-authors, for the judicial, legislative and political relevance of the research on implicit bias.

22 This holds, at least for Holroyd (2011; 2012:303, note 15). Madva (2012:98, note 3), too, is a compatibilist. I suspect that Washington and Kelly (see this chapter, note 19) are also compatibilists.

responsibility and the notion of implicit bias.²³ As both notions presuppose a distinction between the conscious and the unconscious – responsibility connecting to the first term while implicit bias is linked to the second term –, they share the vocabulary of the ‘conscious’. It is then possible to define the problem of responsibility for implicit bias by focussing only on the epistemological interpretation of the conscious/unconscious distinction.

It is possible also to provide an *ontological* interpretation of implicit biases. Research on implicit bias is part of a broader research programme within social psychology that takes the *new unconscious* (or sometimes: automaticity) as its key notion.²⁴ It has replaced the Freudian, psychoanalytic unconscious (also see chapter 3, section 2) and may also be described as a ‘behavioural unconscious’ or ‘natural unconscious’: ‘Unlike the psychoanalytic unconscious, it has no innate drives that seek gratification without regard to constraints of reality and society. In fact it is rather cold, apparently rational, and amotivational, compared to the heat and irrationality of psychoanalytic drives and conflicts’ (Uleman 2005:5).²⁵ Introducing this notion has been a move beyond mere behaviourism, an older position among psychologists who favoured a naturalistic outlook.²⁶ It forges a theoretical link between merely neurological processes and automatic yet socially meaningful human actions without the need to rely on what subjects themselves, during psychological experiments, report on what their actions mean. It has therefore helped social psychologists to explain human behaviour as the operation of internal, automated, but unconscious goal structures.²⁷ Implicit biases, too, belong to these unconscious goal structures. For this research programme, there is no reason in advance to assume that the unconscious workings of the mind are limited to implicit bias. Implicit biases are only one

23 To deal with the problem of (ir)responsibility for implicit bias, several authors choose to shift towards *attributive* (or: *attributionist*) notions of responsibility (see, e.g. Brownstein 2015; Zheng 2016; for more references: see Brownstein 2015:2, note 8). The problem with this strategy, however, is that these authors seem to waver between two different vocabularies for describing human agency: one vocabulary that describes the facts of mind in terms of the conscious/unconscious distinction and another that understands this discussion in terms of the ‘attribution’ of (moral) responsibility. Choosing the first vocabulary allows theorists to employ the notion of implicit biases in describing hidden injustices in society, but simultaneously commits them to acknowledge the problem of responsibility for implicit bias. Choosing the second strategy solves the problem of responsibility for implicit bias but also takes away the possibility of describing hidden injustices. The challenge, of course, is to find a vocabulary that performs both functions: describing hidden prejudices in society (in terms of implicit biases or other terms) and holding people responsible for them.

24 In section 5 of this chapter, I will say more about the uses of terms like ‘unconscious’ and ‘automaticity’.

25 A similar description of the differences between the Freudian and the ‘new’, i.e. the cognitive, unconscious is provided by Kihlstrom, Barnhardt and Tataryn (1991:789): ‘Their unconscious [i.e. of Freud and his psychoanalytic colleagues] was hot and wet; it seethed with lust and anger; it was hallucinatory, primitive, and irrational. The unconscious of contemporary psychology is kinder and gentler than that and more reality bound and rational, even if it is not entirely cold and dry’. For some influential articles and anthologies on ‘the new unconscious’, see: Kihlstrom (1987), Uleman and Bargh (1989); Hassin et al. (2005), Bargh and Morsella (2008).

26 Bargh and Ferguson (2000).

27 See, e.g. Bargh and Ferguson.

category of unconscious attitudes next to many others. It means that many other actions that people consider theirs may be unconscious as well.

Social psychologists have usually ignored the issue of free will as one that belongs to the expertise of philosophers, not theirs.²⁸ But in recent decades, some of them have taken a keen interest in the debates by philosophers of mind, including their discussions on free will as a metaphysical notion.²⁹ For them, the notion of the new unconscious does more than just explain people's unconscious behaviour. It functions as the focal point of a new ambition in cognitive and social psychology: to reduce or even eliminate the consciousness as an explanatory notion by exploring ways of explaining human (social) actions based on the new unconscious.³⁰ For these psychologists, as Kihlstrom observes:

'it seems that automaticity is the key to the scientific status of psychology itself. Automaticity does more than demystify unconscious mental life: It permits us to bypass the will [...] and allow psychology to adopt the pinball determinism of classical physics. Bargh, Wegner, and others, faced with an apparent conflict between free will and determinism, choose determinism' (2008:173).³¹

To achieve their ambition, experimental psychologists use for their research the new and improved instruments and techniques neuroscience has to offer, such as functional magnetic resonance imaging (fMRI) measures. These have helped them to refine their investigations and to explore new domains of human behaviour, including those that are traditionally thought to be the result of conscious deliberation. It enabled them to devise experiments, as some claim, that show people's idea of free will to be an illusionary one.³² Psychologists who are committed to the research programme based on the 'new unconscious' have therefore concluded that it 'is not necessary to invoke the idea of free will or a nondetermined version of consciousness as a causal explanatory mechanism in accounting for higher mental processes in humans' (Bargh and Ferguson 2000:939). And so, as experimental psychology progresses, its techniques are refined and its theoretical framework improved, the unconscious, like a blind cold Leviathan, tends to devour more and more what previously was explained in terms of the conscious. Society, as it turns out, is populated with automatons. The results of ever more psychological experiments have inspired both psychologists and philosophers to make a case against human agency and responsibility and to argue that the unconscious – the ontological

28 Roediger, Goode and Zaromb (2008).

29 E.g. Wegner (2000); Bargh and Ferguson (2000). Also, see several anthologies: Pockett, Banks and Gallagher (2006); Baer, Kaufman and Baumeister (2008); Sinnott-Armstrong (2014).

30 Greenwald (1992); Greenwald and Krieger (2006).

31 Kihlstrom specifically mentions Wegner (2002) and Bargh and Ferguson (2000).

32 Libet (1985).

dimension of the mental – shows the failures of our self-consciousness as self-knowledge – the epistemological dimension of the mental.³³

The current debate on responsibility for implicit bias ignores the Determinism Thesis. This is a mistake, I think. The research programme of the new unconscious that underlies research on implicit bias offers both an epistemological and ontological or, as we may also call it, a ‘neurological’ interpretation of the conscious/unconscious distinction.³⁴ Within this programme, they cannot be taken apart. According to a neurological interpretation, as we have seen, practices that involve implicit biases can be understood as cases of deterministic practices. If participants in the current debate on responsibility for implicit bias choose to be compatibilists, they are adopting a compatibilist stance towards human (social) actions while adopting a (hard) determinist stance only towards actions that originate in implicit biases. This is, of course, an inconsistent strategy. Instead, they should recognise the deterministic challenge that cases of implicit bias raise and generalise this challenge to all human actions. Viewing the problem of responsibility as a local problem, limited only to implicit bias, not as a generalised problem, is mistaken. We cannot use, for the assessment of responsibility, some of the research results that cognitive and social psychology produce and ignore their deeper, naturalistic commitments.

4.3. The manipulation argument

Is Raskolnikov responsible for killing Alyona Ivanovna? Are the Czech neo-Nazis responsible for killing the anonymous, homeless Romani woman? When Alyona Ivanovna was murdered, the act of killing was Raskolnikov’s. It may be regarded as one that was perfectly within his *reflective control*. It has been suggested, however, that during the killing, Raskolnikov was in a state of hypnosis.³⁵ It provides him with a possible defence against conviction for murder. In a hypothetical discussion of Raskolnikov’s case, Bergelson explores whether this defence would hold based on the Model Penal Code, in American jurisprudence ‘the most systematic and widely adopted theory of criminal justice’ (1996–1997:921). In the end, her answer is negative. She observes that Raskolnikov’s situation before the act of murder took place displays several ‘elements of conscious decision-making’, showing that ‘he was still not completely “programmed.”’, and therefore concludes that despite his hypnosis ‘he did not completely lose the ability to control his

33 E.g. Bargh (2008); Bargh and Earp (2009); King and Carruthers (2012); Wilson (2002); Caruso (2012); Kornblith (2012); Doris (2015). Also see several anthologies: Pockett, Banks and Gallagher (2006); Baumeister, Mele and Vohs (2010);

34 In their strategy of explaining the mind, the naturalists’ starting point has, of course, been to take the biological facts of the human body as somehow ‘causing’ the facts of mind. This is, generally speaking, the biological interpretation. The neurological interpretation is one that, in explaining the mind, privileges the neurological facts of the human body.

35 Nuttall (1978).

conduct' (1996–1997:925–926).³⁶ Raskolnikov is responsible. Hard determinists would disagree. They deny that the faculty people have for practical rationality is within their *reflective control*. From their viewpoint, freedom is a mere illusion. As they deny freedom, they also deny responsibility. Alyona Ivanovna has died, the Romani woman has died, nobody is responsible.

What both soft and hard determinist generally agree on is that responsibility presupposes freedom.³⁷ For that reason, the debate usually centres on freedom as a condition of moral responsibility, rather than on moral responsibility itself, despite the many articles and books that have 'moral responsibility' in their title.³⁸ Determinists also agree that responsibility and freedom are intimately linked: the standard option in their debate is between recognising both freedom and responsibility, i.e. as compatible with determinism, or denying both.³⁹ For hard determinists, the obvious aim of their strategies is to deny freedom. But to make an argument against the *existence* of freedom, at least they need a *notion* of freedom, one they agree on with soft determinists.

Both soft and hard determinists usually work with two different types of notions: free will and free action. The second type treats freedom as a dispositional concept: freedom is basically viewed as the ability people have to control their actions (Fischer 1982). This has usually been interpreted as having the ability to do otherwise. Central to this debate is, therefore, the exploration of the alternate possibilities principle.⁴⁰ People are therefore free if they could have done otherwise. The first type takes freedom to be a volitional concept. The analysis of freedom centres on the *formation of free will*: what inner process will guarantee the formation of a will that deserves to be called 'free'?⁴¹ People have free will if they are somehow able to interrupt the stream of causes, prevent (some) previous causes from influencing their will and instead establish themselves as the 'source' (or: the 'author') of their will: their will is their *own*. We might call this the

36 One of the sources for her argument about 'conscious decision-making' is the Model Penal Code, § 2.02(2)(a)(i): 'A person acts purposely with respect to a material element of an offense when [...] if the element involves the nature of his conduct or a result there of, it is his conscious object to engage in conduct of that nature or to cause such a result'.

37 For a more nuanced explanation of what this means, see Zimmerman (1997:411), who calls freedom a 'root requirement of responsibility'. Some determinists are not explicit on this to the point of even conflating moral responsibility and freedom. But when they claim to argue against moral responsibility, it is really freedom they try to undermine.

38 In her critical analysis of what she calls the 'modern concept of moral responsibility' Smiley already observed, and not without a sense of irony: 'Note in this context the number of volumes entitled Moral Responsibility and [...] which are in fact about free will and determinism' (1992:82). Ever since Smiley published her work, many volumes with this sort of title have continued to appear in abundance.

39 One notable exception is Waller (1990), who argues, against compatibilists, that determinism is incompatible with moral responsibility and, against hard determinists, that it is compatible with freedom.

40 See Widerker and McKenna 2003.

41 The problem of the formation of free will has also been called the problem of the 'origins of one's will' (Watson 1987:151), the 'source' (McKenna and Coates 2015), 'ultimately responsibility' (Kane 1996).

‘ownership’ claim about freedom.⁴² It is common to almost all definitions of the first type, whatever their many other differences. To conclude: freedom does not just consist in people’s control over their actions, but in their control over their actions based on a will that is their *own*.⁴³

My discussion in this section will be limited to the notion of freedom belonging to the first type: the formation of free will.⁴⁴ One of the main arguments in favour of free will is provided by the notion of reflective volitional autonomy, which sustains the claim that consciousness provides people control over their actions. Many hard determinists who try to argue against this argument mention Frankfurt’s account of freedom (1971) as representative of this argument. It was discussed in the previous chapter. Some of them have denied consciousness altogether. Others have been willing to concede that consciousness exists only to deny that it would make any difference.⁴⁵ These last ones understand nature as putting limits to introspection: they question to what extent, if at all, people are able to achieve, based on their critical introspection, control over their will and their actions. In this section, I will address in particular one influential argument that the second group of hard determinists have used against the liberal notion of freedom, and which is called the Manipulation Argument.⁴⁶ It will be discussed in an authoritative version presented by Pereboom, a prominent theorist in support of hard determinism or, as he once called it, ‘determinism *al dente*’.⁴⁷ His version is especially interesting because he explicitly argues against Frankfurt’s hierarchical account of free will and therefore against reflective volitional autonomy.⁴⁸ It has been presented in

42 Wolf (1990, chapter 2) calls it ‘the Real Self View’.

43 Determinists usually argue about both types of definitions in distinct debates. I have found little literature explaining the links between both types. Watson (1987) at least seems to recognise them as two features of supposedly one notion of freedom. Nevertheless, he is unclear about this: he labels both types as respectively ‘free will’ and ‘free action’, but it obfuscates that both types have a common hinge: the will. Both types also share a criterion of what makes a will or action free: the proper kind of control, either internal (‘free will’) or external (‘free action’).

44 I intend to argue that the liberal notion of responsibility cannot stand up to naturalist (determinist) criticisms. In the next chapters, I will develop an alternative notion that will. Addressing only the first type and leaving aside the second type may yet render the proposed alternative vulnerable to determinist criticisms of the second type. I do not think this will be the case because in developing my alternative account, I will also raise doubts about the determinist thesis itself. Apart from that, the debate on the second type seems to have put hard determinists in the position of defending their position. Decisive was an article by Frankfurt, ‘Alternate possibilities and moral responsibility’ (1969), whose main point was that the ability to do otherwise is no necessary condition of moral responsibility. As Widerker and McKenna (2003:6) put it: ‘Frankfurt’s argument, if sound, challenges the incompatibilists and libertarians to explain why, in a deterministic universe, no agent is morally responsible’.

45 Here I mention ‘consciousness’ as what these philosophers concede, but depending on the type of argument or on the specific feature of consciousness they address, they may also refer to, for example, ‘introspection’, ‘first-person perspective’ or ‘self-knowledge’.

46 See e.g. Levy and McKenna (2009:107-111); McKenna (2012); Tognazzini (2014); Fischer (2016); Mickelson (2017).

47 Pereboom (1995).

48 Pereboom addresses Frankfurt’s account of free will in at least three of his writings (1995; 2001; 2013). I will take my lead from (2001), but I will also mention some differences.

several of Pereboom's writings, but in each of them, the strategy is the same: Pereboom introduces a story about a man killing a woman and then presents four varieties of how the murderer is manipulated into committing the murder. The manipulation cases are part of an argument for hard determinism.

Pereboom starts telling a story that compatibilists could recognise as a *standard case of moral responsibility*. He introduces the character of Professor Plum, who will remain the protagonist in the four cases yet to come. The basic storyline is simple: Professor Plum is acquainted with Ms White and decides to kill her for personal advantage. Pereboom does not provide any details on their relationship. To make the story somewhat more appealing, we may follow Garvey's example in adding some more context (this will not affect Pereboom's argument):

'Ms. White has named Professor Plum as the sole beneficiary of her life insurance policy. Finding himself in need of cash to supplement his sorry salary he decides to kill Ms. White and reap the benefits.' (Garvey 2015:48)

How should we understand Plum's motivation to kill Ms White? First of all, Plum has an egoistic character. For him, egoistic reasons usually 'weigh very heavily – much too heavily as judged from the moral point of view' (2001:111). This way of explaining Plum's motivation would allow Plum to be a wanton, in Frankfurt's view, and therefore make him unfit for the leading role in a standard case of moral responsibility. But Pereboom makes clear that we should understand Plum's motivation in terms of second-order volitions:

'his desire to kill White conforms to his second-order desires in the sense that he wills to kill and wants to will to kill, and he wills to kill because he wants to will to kill.' (Pereboom 2001:111)

So Pereboom uses the reflective structure of motivation, as explained by Frankfurt, to sketch a standard case of moral responsibility.

As Pereboom intends to criticise several (pairs of) compatibilists theorists – apart from Frankfurt (1971) also Wallace (1994), Fischer and Ravizza (1998), Hume and Ayer (1954) –, he endows Plum's motivational structure with features these theorists deem necessary for freedom. It allows him to show that even these features provide no warrant for Plum, or anyone else, for being free in a deterministic world. I will not discuss them except for one which states that Plum (2001:111):

'is not completely egoistic, and indeed he retains the general capacity to grasp, apply, and regulate his behavior by moral reasons. For example, when the egoistic reasons that count against acting morally are relatively weak, he

will typically regulate his behavior by moral reasons instead. These capacities even provide him with the ability to revise and develop his moral character over time.’

Here Pereboom prepares the ground for an argument against Wallace, who defends the view that a person can only be held responsible if he possesses at least the powers ‘to grasp and apply moral reasons’ and ‘to control or regulate his behavior by the light of such reasons’ (1994:157).⁴⁹ For now, I will call them ‘moral powers’. Even if Pereboom had not chosen to argue against Wallace, he still should have ascribed these powers to Plum. The story about Plum is an example of what Dennett would call ‘intuition pumps’: thought experiments that are used ‘to expose what we take to be conceptual problems in the views we expose’ [1984] (2015:191). Dennett uses them himself, but he also warns for their abuse [1984] (2015:13):

‘Perhaps the most frequent abuse is deriving a result – a heartfelt intuitive judgment – from the very simplicity of the imagined case, rather than from the actual content of the example portrayed so simply and clearly.’

Pereboom would agree with this.⁵⁰ Plum should therefore appeal to people’s intuition that he is basically a responsible character. It is therefore important for Pereboom to portray Plum as a being that is sufficiently complex. If Plum would be a consistent egoist, we would [people might] view him, in his relations with other people at least, as a sort of ‘human automaton’, a ‘one-dimensional man’, or perhaps: a ‘moral idiot, a sociopath’ (Oshana 2002:269). Would Plum still appeal to our intuition that we deal here with a person that we regard as free? Wallace claims that being free *requires* ‘being moral’, in the sense of having the ability to adopt a moral stance, i.e. of having the two moral powers already mentioned. He understands them as ‘powers of reflective self-control’ (1994:157). Although having these powers is no guarantee against sometimes making egoistic choices, at least a consistent egoist would not be free in this view.⁵¹ He would be a psychopath or, we might say, a *moral wanton*. Frankfurt could, of course, maintain that people are free when their first-order volitions are determined by their second-order volitions, even if these second-order volitions display a consistent egoistic

49 See in particular Wallace (1994). In a comment on Frankfurt’s account of freedom, Wallace himself remarks that Frankfurt is ‘interested primarily in the problem of autonomy rather than that of responsibility’ (1994:14-15).

50 I am not convinced that Pereboom, in the end, manages to paint a persuasive picture of Plum, simply because the features he borrows from several compatibilists and ascribes to Plum do not, in my opinion, result in a coherent rational-psychological picture of Plum. But this does not affect my discussion of Pereboom. For my own purposes, I assume that at least the features Frankfurt and Wallace believe are important are in harmony.

51 Cf. Wallace’s comment on Kant: ‘it is possible that agents might have the general powers of self-control without having strong freedom of the will’ (1994:14).

pattern.⁵² In fact, he remains silent on the issue of egoism. In a passing remark, he denies ‘that a person’s second-order volitions necessarily manifest a *moral stance*’ (1971:13, note 6), but this still allows his recognition that the ability to adopt a moral stance may also be a necessary condition for freedom. In the end, the accounts of Frankfurt and Wallace can be viewed as complementary. As Wallace describes the moral powers as ‘powers of reflective self-control’ (1994:157), his account too relies on the idea that the practical mind has a reflective structure. While Frankfurt spells out what this reflective structure amounts to, Wallace explores the meaning of having these moral powers.⁵³ Because Pereboom includes Wallace’s account into his story on Plum and uses it to make Plum into a more complex character, this adds to a more persuasive story.

In the next part of his argument, Pereboom presents four cases as varieties of this basic story. The fourth and last case simply presents Professor Plum in a world that is taken to be deterministic and therefore represents the world as it is, according to Pereboom. The first three cases tell the story of Professor Plum as well, but in a world that is presumed to be non-deterministic, a world, I take it, that is basically compatible with our intuition that we are free. But even in such a world, it is possible, as many will agree, that people are somehow deceived or manipulated. This is what Pereboom intends to present: three different examples of manipulation. The first two examples seem to be taken from the science fiction genre. Pereboom does not want us to think that the science-fictional cases may be true one day. Again, as with Plum’s character, Pereboom only wants the three cases to appeal to people’s intuition. People should think of them as genuine cases of manipulation. I will briefly discuss the three cases.

The first and second cases have in common that Pereboom presents neuroscientists to perform the role of manipulators. In both cases, Pereboom attributes to them the power to have created Plum. He does not comment on what this exactly means. Does he think of them as Victor Frankenstein, the scientist that created his own monster? It is nevertheless clear that having created Plum provides them with the opportunity of manipulating him to an excessive extent, in the first case ‘through the use of radio like technology’, in the second case by programming him.

Pereboom’s choice for neuroscientists as fulfilling the role of manipulative evildoers – for they are, in cases one and two, responsible for the murder – fits in a tradition that uses the character of a devilish being as part of a philosophical argument and which

52 Cf. Wallace’s discussion on the egoist and his (in)capacity for moral sentiments, such as resentment (1994:248).

53 Wallace acknowledges that both he and Frankfurt intend to analyse ‘practical freedom’, but, as he understands their differences between their approaches, Frankfurt is interested ‘primarily in the problem of autonomy rather than that of responsibility’ and in the question of ‘whether and how we exercise the powers of reflective self-control’ rather than ‘whether we possess’ them (1994:14-15, note 22). The idea that autonomy and moral responsibility are only different aspects of the same phenomenon, i.e. ‘practical freedom’, seems to be more or less widely, although mostly tacitly, recognised. It is disputed, however, by Oshana (2002).

is at least as old as early modern philosophy.⁵⁴ In the second half of the twentieth century, the role of the devil is taken over by the figure of the neurologist, probably first introduced by Richard Taylor (1963). John Martin Fischer (1982) reintroduced the devilish element by adding the adjective ‘demonic’ in the phrase. The idea of what neurologists (in some examples also: neurosurgeons or mad scientists) do with people (‘rewiring’, ‘inserting beliefs’, ‘reprogramming’) elicited from Dennett the brief remark ‘still fortunately science fiction’ (1973:175). Unfortunately, the figure of a demonic neurologist who manipulates people, although not exactly in the sense that Dennett had in mind, is not as science-fictional as determinists may think. Under the Nazi regime in Germany, a neurologist, Robert Ritter, contributed to the genocide of Romani people through his racial-scientific research on Romani people.⁵⁵ Again, we have no need for fictional examples.

In the third case, the manipulation by neuroscientists is replaced by manipulation performed by a community. Plum has been born and raised in a community of people that has given him ‘rigorous training practices’ with the result that he is ‘a rational egoist’ (1995:24). Here we arrive in a world that seems closest to the world we believe we live in. We may think of this community perhaps as a sect that subjects its members to brainwashing. Should we cherish any hope that Plum has the ability to somehow remind himself of other people’s humanity as of his own, and to develop into a substantially less egoistic character’, Pereboom makes clear that ‘his training took place at too early an age for him to have had the ability to prevent or alter the practices that determined his character’ (2001:114). This does not mean that Plum, as we discussed already, is exclusively rationally egoistic. He is not. His egoistic reasons may be very powerful, but the desire to which they give rise is not irresistible.

In the fourth case, the manipulation of Professor Plum is performed, speaking metaphorically only, by the natural forces themselves who are disenchanting, blind and non-rational and perform their manipulations in a non-intentional, mechanical way only. Compatibilists and hard determinists agree on the description of Plum’s murder provided in this case: the setting is a deterministic world that is populated with certain beings, among them Professor Plum, who have certain features that compatibilists think are special. But they disagree on what these features imply. Compatibilists believe that these provide the key to understanding how Plum can be held responsible for the killing because they help us to understand how certain causes for the desire to kill (and therefore for the killing), i.e. the egoistic reasons, are Plum’s *own* reasons and therefore as what makes Plum free in a deterministic world. Here Plum uses the first three cases to argue that compatibilists are mistaken. His basic argument runs like this. The first three

54 Perhaps the first one to use this character within the context of a discussion on free will was John Wisdom (1934), who stages in one of his examples a devil that fixes the strength of people’s various desires whenever they make a decision. Of course, Pierre-Simon Laplace, at the start of the nineteenth century, introduced already the figure of a demon: but see this chapter, note 17.

55 Schmidt-Degenhard (2008).

cases illustrate intuitively that Plum, despite his ‘special’ features, is not responsible for killing White. The explanation for this would seem to be that his desire to kill White is traceable to ‘factors beyond his control’ (2001:116): the interventions by neuroscientists in the first and second cases, the rigorous training practices by his community in the third case. If we agree with this explanation, then in the fourth case, we should conclude that Plum, despite his ‘special’ features, is not responsible for killing White because here too, his desire to kill White is traceable to ‘factors beyond his control’.

The first three cases postulate the presence of certain agents who are in fact responsible for producing in Plum the desire to kill White. Compatibilists might argue that this difference is crucial for their argument. Pereboom denies this. But this is not crucial to the Manipulation Argument. In each of the four cases, a certain causal process is identified as ‘manipulation’, ranging from a process that takes only a short time, as in the first case, to a process that, in the fourth case, last indefinitely, comprising all of nature’s chain of causes. What matters is not how this process is *produced* (by people, by a machine or by nature), or how long it takes, but what it *produces*: *egoistic reasons* that are forceful enough to spark in Plum a desire to kill Ms White. And whatever powers Plum has been endowed with by compatibilists, including his power for critical reflection, these do not help him to gain control over his egoistic reasons.

4.4. The reflective belief challenge

In the preceding section, the Manipulation Argument was discussed as an argument that undermines reflective *volitional* autonomy (an argument *for* freedom) and thus aims *against* freedom. In this section, I want to discuss a case showing the relevance of the link between ‘will’ and ‘belief’. It allows me to introduce reflective *epistemic* autonomy (already discussed in chapter 2, section 5). It possibly counters the Manipulation Argument. It means that the scenery for the determinism debate will be changed. We will move from the issue of people’s *control over their will*, which presents the standard approach in the determinist debate on freedom, to the issue of people’s *control over their beliefs*. First, I will explain the relevance of linking these issues for the debate on responsibility for one’s prejudices, then discuss a case that links prejudice and ‘control over belief’ and use this to interpret the Manipulation argument as claiming that we have no control over our beliefs. I will conclude with a brief discussion of the debate on ‘doxastic voluntarism’ as the relevant context for reflective epistemic autonomy.

Will and belief are connected in a way that is relevant to understanding people’s responsibility for their prejudices. This is illustrated by a case that Garvey (2008) uses to analyse just this connection.⁵⁶ In 1984 a man boarded a New York subway. He was

56 Garvey, the author of this article, analyses the connection from the specific scope of holding people responsible for their prejudices in the context of penal law. As I am concerned here with an ontological-moral context, it is beyond my present scope to discuss the differences between Garvey’s and my approach.

approached by two of four young men who had been ‘noisy and boisterous’, one of them asking him: ‘Give me five dollars’ and then asking again. As the man felt threatened and believed he was about to be killed, he drew his gun and fired five shots at the four men: they survived but were nevertheless wounded seriously. The man, however, was mistaken. None of the four men had carried any gun. As the shooter was charged with attempted second-degree murder, he claimed he had acted in self-defence.-

Legal or moral cases are usually analysed as a problem of people’s responsibility for action. It requires that we identify in the actor some ‘will’ (‘intent’, ‘choice’, ‘decision’ or, in the proper juridical vocabulary: *mens rea*, a culpable state of mind) and a causal relation between this ‘will’ and the action. Garvey is interested in analysing cases as a problem of people’s responsibility for their beliefs.⁵⁷ The case of the subway shooting allows him to address this problem because invoking self-defence, as the subway shooter did, renders the question of responsibility for action irrelevant. It is taken for granted that the actor is responsible for his action. The subway shooter did fire five shots. Instead, the legal or moral analysis is shifted towards the question of whether the actor could be excused for his action. To answer this question, Garvey understands people’s responsibility for action as somehow depending on the beliefs they have. The case of the subway shooting then raises the question of whether or to what extent the shooter is responsible for all the beliefs he had when he pulled the trigger, and which together were sufficient, at that moment, for him to form the ‘will’ to shoot. The belief that immediately gave rise to this will, it may be claimed, was the belief that *p*, i.e.: ‘I am about to be killed, and deadly force is necessary to avoid being killed’ (2008:124). It expresses the believer’s expectation about what will happen to him and a practical judgement about what he, therefore, ought to do. It is possible to trace the belief that *p* back to what Garvey calls ‘background beliefs’, which may include, for example:

‘beliefs related to the victim’s movements or gestures, his request or demand for five dollars, the tone of his voice, the look in his eye, the tight confines of the subway car, and so forth.’ (2008:141)

If only these beliefs are assumed to be sufficient to form the belief that *p*, they would suggest an excuse for the shooter: he had these beliefs, he had no control over these beliefs, and therefore forming the intention to shoot was inevitable.

Garvey did not select just any case of self-defence. The subway case has a twist to it. The case is drawn from American jurisprudence and is known as the Goetz case, named after the shooter. Goetz was a white man. The four men wounded by him were black. At

57 I take some liberty in presenting Garvey’s approach to the problem of self-defence. For example, he describes the problem in terms of the reasonableness of the actor’s beliefs rather than the actor’s responsibility for his beliefs. I choose to rephrase it in terms of responsibility to keep the vocabulary of my own argument consistent. This does not, I think, provide a distorted presentation of Garvey’s argument. Garvey addresses the ‘morality of responsibility’ in yet another article (2015).

the time, the killing provoked much public discussion. According to some:

‘Goetz believed that he was about to be attacked because he was a racist, and would not have believed that he was about to be attacked had he not been a racist.’ (2008:124, note 27)

This claim turns our attention to another class of beliefs involved in the actor’s decision to shoot: not only Goetz’s perceptual beliefs, i.e. beliefs he formed based on what he perceived in that situation but also his social beliefs matter to assess excusability, in this case, the generalisation that ‘blacks are more prone to violence than nonblacks’ (2008:141).⁵⁸ This is how Garvey analyses the gist of the problem:

‘Call this belief the belief that *q*. I will assume that this belief was necessary to [Goetz’s] formation of the belief that *p*, and that it, together with his other background beliefs, were sufficient to cause him to form the belief that *p*. [Goetz’s] possession of the belief that *q* is usually what leads to his characterization as a racist, and inasmuch as his racism consisted in his possession of that belief, [Goetz’s] racism was cognitive.’ (2008:141)⁵⁹

The issue of excusability appears to turn on the question of whether Goetz can be said to be responsible for his racist belief.⁶⁰

Garvey’s presentation of the Goetz case has relevance beyond the legal discussions he himself is concerned with because it opens a new line of thinking for the determinist debate on freedom. Perhaps we need to accept the idea, as determinists suggest, that people’s will is itself determined by prior events. However, if some of these events, such as beliefs, if only a subclass of them, can be understood as falling within the range of people’s responsibility, then this opens the way for a different claim of people’s freedom, one that traces people’s free will formation to their free belief formation. This is also how we may look at Frankfurt’s account. Even though Frankfurt chooses to address the problem of freedom as one of free will formation, his account presupposes the possibility

58 Garvey mentions some other generalisations: ‘We can also assume that [the shooter’s] background beliefs included two other beliefs: that males are more prone to violence than females; and that the young are more prone to violence than the old. These beliefs are of course generalizations’ as Garvey remarks: ‘Even so, I doubt that many people would consider them illegitimate bases upon which one might form the belief that *p*.’ (2008:141).

59 Actually, Garvey has, for the purposes of his argument, transformed the case of Goetz into the case of what he calls ‘Goetz*’, which includes some adaptations and simplifications of the original Goetz case. But, as he remarks: ‘I doubt that the features I strip away from the real case are really what matters to most people who believe that the real Goetz did not act in self-defense’ (2008:124, note 27). I agree with him and refrain from discussing the differences between both cases.

60 I will not further discuss Garvey’s analysis of the case. Although Garvey refers to both philosophical and psychological arguments, obviously, he takes a legal interest in discussing the case of the subway shooting. But a purely legal discussion does not exhaust the moral aspects of epistemic responsibility.

of belief formation, more specifically of *introspective* belief formation, i.e. self-knowledge as having knowledge of one's mental states.⁶¹ The issue of free agency may therefore be reconstructed as an issue of belief formation. Before I proceed to explore this possibility, I will briefly return to the Manipulation Argument and show its limitation from the present discussion.

The Manipulation Argument is a Trojan Horse. It carries within its belly the assumption that people have no control over their beliefs. This is implied by the use of manipulation as a device to construct thought experiments. To explain this, I will once again look at the first three cases and ask how the manipulation in these cases is supposed to work.

In the first place, Pereboom does not, in the first three cases, understand the power of manipulation in an absolute sense. It does not give their owners, for example, direct control over Plum's body. This would have provided the most effective means, for example, for the neuroscientists, of realising their goal: killing Ms White. It is clear from the examples, however, that Plum's mind is being manipulated, not his body. But even in manipulating his mind, the power of manipulation again does not extend that far that its owners are able to directly create the mental state that causes the act of killing (whatever we call this mental state: either the desire, the will or the intention to kill). The reason for putting limits on the power of manipulation is perhaps, again, to make the cases appealing to people's intuition. If the manipulators were granted some godly power – the power of direct intervention –, the metaphor of manipulation would probably lose its credibility.

The ways of exerting control over Plum's mind, or parts of it, vary among the three cases, but in each case, the object of manipulation remains the same: the reasoning process, because in the first case, Plum is manipulated 'to undertake the process of reasoning by which his desires are brought about and modified' (2001:112–113), in the second case he is 'programmed [...] to weigh reasons for action' (2001:113), again with the effect of bringing about his desires, finally in the third case he is (case 2), or because he is trained 'in such a way that his reasoning processes are often but not exclusively egoistic' (2013:426).⁶² The different forms of manipulation are effective because in each case, the reasoning process is controlled and diverted in the direction of certain reasons, in this case, egoistic reasons, just enough at least to tip the balance unfavourably for the moral reasons.⁶³ But while the acts of manipulation have as their object Plum's reasoning process, rather than his desires, again they do not consist in controlling the reasoning process directly, in any absolute

61 Frankish (2004:159) interprets Frankfurt's account in this way.

62 Pereboom's phrases remain fairly consistent throughout his (1995), (2001) and (2013) for all cases, with the exception perhaps of the third case. Here the fact that manipulation targets the reasoning process is made explicit only in (2013).

63 What is relevant here is not that these reasons are egoistic. It is also possible to construe the cases in a way that the manipulation makes him choose social or moral reasons (which would not have ended in killing Ms White).

sense. Whatever power the neuroscientists have, Pereboom does not lend them the ability to shape the reasoning process itself. This becomes clear when Pereboom explains that, in the first case, the neuroscientists manipulate Plum by ‘pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic’ (2001:113). So, manipulation is a matter of exerting influence, of having control only in a mediated way.

In all three cases, manipulation consists in exerting influence on the reasoning process.⁶⁴ If my reading of Pereboom is correct, then how exactly should we understand this? Pereboom does not provide any specifics. What we do know is that the reasoning process should produce, in some way or another, certain kinds of reasons: egoistic reasons. So how should the reasoning process look like that produces such reasons? It may be helpful to take in mind Dostoyevsky’s story on Raskolnikov, the murder of the anonymous Romani woman in Prague or even the Goetz case. Each of the murders had his reasons to perform the murder they did. What helped them in forming these reasons were certain stereotypical beliefs about society, about rich people, about women, about Romani, about black people, about carrying guns. Each of them took these beliefs, *their* beliefs, for a fact about the society they lived in. My suggestion is that we understand the reasoning process people engage in as being embedded within, and taking place before, a background of beliefs. It is not possible to imagine people engaging in a reasoning process, whatever else it consists in, without also having certain (and probably: a lot of) beliefs. This suggestion provides at least one possible way of understanding what manipulation in Pereboom’s cases means: manipulation consists in the creation of mental states, best described as ‘beliefs’, that push the reasoning process in the direction of certain reasons.⁶⁵ For precisely this reason, the manipulated beliefs can be forceful enough to make Plum form the intention of killing Ms White. And whenever these beliefs are not forceful enough, the neuroscientists may produce new beliefs, as many and as subtle as are needed to induce in Plum the rise of egoistic reasons.

If we stop our analysis of manipulation here, then we have not yet grasped why the acts of manipulation must be effective. Pereboom needs to assume that Plum is unable either to reflect on the ‘origins’ of his beliefs or to change his beliefs at will. If he did not make this assumption, if he did not exclude that Plum has mental abilities to change his beliefs at will, Plum might yet escape his manipulation.

This brings us again to the central issue of the present section: do people have control over their beliefs? By now, it should be clear that the Manipulation Argument not only tacitly assumes but even owes its force to the assumption that people have no control over their beliefs. Whoever accepts the manipulation metaphor and the

64 I take it that Pereboom understands the reasoning process as included in the reflective process.

65 Another suggestion might be that manipulation consists in undermining the reasoning process itself, i.e. rendering this process somehow irrational. But the cases provide no reason, I think, to assume that this is a plausible suggestion.

argument it supports is also committed to this tacit assumption. Within the determinist debate on free will, the possibility of control over beliefs has not usually been taken into consideration, nor even the relevance been recognised of the link between will and belief formation. Fischer, one of Pereboom's prominent opponents in the determinist debate, has accepted the force of the manipulation device, as have others involved in the debate, and therefore commits himself to the assumption that people have no control over their beliefs. It seems, as both soft and hard determinists agree on this, they have felt no need or interest to question this assumption. I think this is a mistake. As we now can conclude, the claim of reflective volitional autonomy is defeated only by the Manipulation Argument if we accept the assumption that people have no control over their beliefs. Another view is possible: to understand the free formation of the will as presupposing yet a different sense of freedom, namely the free formation of *belief*.

What are the options – or limitations – to make sense of the idea of the 'free formation of belief'? For this, I will turn to a debate on what is mostly called 'doxastic voluntarism' or sometimes 'ethics of belief'. In some ways, it is reminiscent of the determinism debate. At stake in both of their debates is freedom: determinists argue about the freedom that is connected to desire, free agency, while their colleagues discuss doxastic freedom, the freedom people have in the formation of their beliefs. And in both debates, freedom is also understood in terms of control, either control over actions or control over beliefs. The determinist debate assumes, as we have seen, the truth of the determinist thesis and then asks whether free agency is compatible with it. The doxastic debate, however, starts from quite different premises. The participants in this debate agree that 'beliefs aim at truth'.⁶⁶ The question this claim raises, not the truth of the determinist thesis, is their starting point: what is the epistemic relation between belief and truth in terms of control?

Within the debate on responsibility for beliefs, the principal opposition would seem to be, at first glance, between naturalists and epistemic deontologists.⁶⁷ Naturalists start from a notion that naturalises the process of our belief formation. Even if they are not committed to the determinist thesis, at least they are committed to the mechanical thesis. So, their position provides no escape from the deterministic view on belief formation. Epistemic deontologists, on the other hand, claim that the relation between belief and truth is essentially normative.

Perhaps we may be content with the deontological view on the relation between belief and truth, leaving aside all the intricacies of the arguments they exchange for or against control because this view already poses an alternative to naturalists' views. But this conclusion is too hasty. Epistemic deontologists are divided among themselves into those who claim and those who deny control over belief. Some believe that the deontological relation between belief and truth allows for control over beliefs: people can decide over

66 For discussions of this premise: e.g. Shah (2003), Owens (2003).

67 E.g. Alston (1988).

their beliefs at will.⁶⁸ They are voluntarists. Others deny this. That people have no control over their beliefs seems to be the standard view (sometimes referred to as ‘involuntarism about beliefs’), a view to which naturalists are also committed.⁶⁹ But it is not uncontested.

The deontological view that denies control over beliefs amounts to a different danger: that of rational determinism.⁷⁰ The difference in view between epistemic deontologists who claim and those who deny control over beliefs turns on the interpretation of the obligation to believe. The first claim is that an analogy can be made between the obligation to believe and the obligation to act. As ‘ought implies can’, as the well-known dictum says, the obligation to act implies that people are able to act against the obligation. The obligation to believe is understood in an analogous way: if people ought to believe *p*, perhaps because of some evidence presented to them, they are nevertheless able to believe something different. Their opponents, however, have argued that this is incorrect ‘by trying to assimilate obligation in belief to other kinds of obligation for which the “ought implies can” principle does not hold’ (Shah 2002:436).⁷¹ Assuming that they are right, their view would offer no better prospect of escaping determinism. Even if it provided a different variety of determinism, it still would imply that people have no control over their beliefs and are not able to change them, including their racist beliefs.

If we are to find some view that does offer an escape from determinism (either naturalist or rational), we need to dig a little deeper into this debate and explore especially those views that claim control over belief based on an analogy between believing and acting. At least one possible view is generally ruled out: that we have *direct* control over our beliefs.⁷² An example given by Alston may illustrate the point (1986:196):

‘When I see a truck coming down the street I am hardly at liberty either to believe a truck is coming down the street or refrain from that belief.’

It leaves us with views that claim control over belief based on the indirect, i.e. mediated, belief formation. This is part of the notion of reflective epistemic autonomy that was introduced and explained in chapter 2 (section 5). This notion supports the view that

68 See, e.g. Steup (2000).

69 As Kornblith comments (2012:86): ‘Voluntarism about belief is not a very widely held position’.

70 Rational determinism was the subject of debates in late Antiquity and the early Middle Ages but is ‘largely ignored’ in modern debates on free will and responsibility (Normore 2002). For an exception: Bourke (1938).

71 As authors who work out this view, Shah (2002) mentions Feldman (2000) and Wolterstorff (1997). The analogy between acting and believing is a familiar dividing line within the debate on doxastic voluntarism. It is defended by, e.g. Pettit and Smith (1996); Railton (1997); Scanlon (1998); Velleman (1996). It is criticised by, e.g. Wallace (2001); Austin (2001); Engel (2002) and Wagner (2015).

72 Garvey (2008:158-162), for example, discusses and rejects this type of view under the heading of ‘belief-choice theory’. Engel (2002) discusses such views in terms of ‘believing at will’. Also, see Alston (1988).

people have freedom of agency based on the free (i.e. reflective) formation of their beliefs.⁷³ In the next two sections, I will show how this argument can, in its turn, be undermined by a naturalist view on belief formation.

Does reflective epistemic autonomy imply any limits to the use of Pereboom's cases as intuition pumps? I think it does. The cases reach their limits once the neuroscientists – whenever the subject of their experiment (Plum) comes close to forming an intention that is unfavourable to them – are supposed to have the power to produce in their subject's mind any belief at their will. At the start of this section, we have seen that in each case, the reasoning process is influenced and diverted in the direction of certain reasons, in this case, egoistic reasons, just enough at least to tip the balance unfavourably for the moral reasons. It may be agreed that people have no control over beliefs in the sense of the beliefs people find themselves having, 'beliefs' as psychological facts about themselves, their 'psychic given'. As the Korsgaard-Moran view of belief formation teaches us, however, at least we may have 'control' over the *effects* of these beliefs. People's mental economy may consist of psychological facts that run their own course, but their basic stance is a deliberative one. Distorted untrue beliefs have no irresistible force. Their force may be suspended by people's avowals of other beliefs.

4.5. The manipulation argument revisited

To make an argument against reflective epistemic autonomy, hard determinists have at least two strategies at their disposal. They might argue that reflective belief formation itself, including what Moran calls 'avowal', is subject too to determinism.⁷⁴ Or they might concede this point to Korsgaard and Moran and shift the discussion towards the second type of freedom: free action. They then might argue that, whatever freedom people have to avow some beliefs and disavow others, once their reflections on beliefs and volitions have resulted in the formation of an autonomous will, their will nevertheless fails to be effective. In this section, I will present a revision of the Manipulation Argument by focussing on the second strategy, one that aims not at undermining reflective thought itself but at severing the causal link between reflective thought and the world. In line with the previous section, I approach the debate on 'free action' as an issue of reflective belief rather than of reflective will. I will argue that the Manipulation Argument may counter the Reflective Argument if it analyses introspection as both self-knowledge and consciousness, i.e. if it combines an epistemological view with an ontological view of introspection. This time I will turn to a psychological instead of a philosophical theory: dual-process theory. If I succeed in outlining this strategy, it results in what might be

73 The debate of doxastic voluntarism offers yet another type of argument: reason-responsiveness, which seems to be a counterpart of the argument on reason-responsiveness in the deterministic debate: see McKenna and Coates (2015, section 5.4) on 'Reasons-Responsive Compatibilism'.

74 This view seems to be advocated by, e.g. Bargh and Ferguson (2000) and Kornblith (2012).

called the *Manipulation Argument Revisited*. Along the way, I will suggest how it can also be used to support the first strategy.

A different type of debate than we discussed so far deals with ‘free action’: it focuses not so much on the relation between an agent’s second-order volition and his will but on the will and his action.⁷⁵ Its central question is: Do people have control over their actions? The standard debate on ‘free action’ has interpreted it as a problem of moral responsibility that involves the Principle of Alternate Possibilities. As it has taken its cue from yet another influential article by Frankfurt (1969), it has rephrased the central question in this specific way: ‘Must an agent be able to do otherwise in order for her to be morally responsible for what she does?’ (McKenna and Widerker 2003:1). This question and the various arguments that have been developed to answer this question will not concern me here because they do not allow me to pursue my strategy to approach freedom as an epistemological problem. The standard debate on ‘free action’ also uses, as does the standard debate on ‘free will’, the metaphor of manipulation to clarify certain points in the argument. And here, too, a variety of the Manipulation Argument can be found. One of its versions again portrays the manipulating villain as a neurologist:

‘Black is a nefarious neurosurgeon. In performing an operation on Jones to remove a brain tumor, Black inserts a mechanism into Jones’s brain which enables Black to monitor and control Jones’s activities. Jones, meanwhile, knows nothing of this. Black exercises this control through a computer which he has programmed so that, among other things, it monitors Jones’s voting behavior. If Jones shows an inclination to decide to vote for Carter, then the computer, through the mechanism in Jones’s brain, intervenes to assure that he actually decides to vote for Reagan and does so vote. But if Jones decides on his own to vote for Reagan, the computer does nothing but continue to monitor – without affecting – the goings-on in Jones’s head’ (Fischer 1982:26).

And this debate, too, presupposes rather than explores that people do not have introspective access to possible manipulations of their consciousness: the acts of manipulation are simply assumed to be effective.

To make sense of ‘free action’ – i.e. control over one’s actions – in epistemological terms, I suggest that it should fulfil the following requirement:

The Monitoring Requirement: Once people’s reflective will formation has resulted in an autonomous will, their execution of this will (i.e. the will to perform a certain action) requires that they are able to monitor whether the action is performed according to plan, more specifically: they are able to monitor and adjust, if necessary, their bodily movements by reflecting on their perceptions.

75 Also, see my remarks in section 2.

This requirement is part of the notion of reflective epistemic autonomy but now transferred to the debate on 'free action'. The monitoring requirement is a necessary condition for the view that people have ownership over their actions, more specific over their physical movements. If it is fulfilled, it excludes some situations from occurring. If, for example, my own will is not effective because my arm remains hanging motionless beside my body, I will observe that something has gone wrong. Although my will is not effective, in monitoring my body, I am able to notice that something has gone wrong with my arm. It would create the opportunity for me to find out what has happened to it. I then may choose, for example, to go to the doctor. Furthermore, I will assume that no time limit is set on monitoring one's own actions. Someone may monitor her actions while performing them. She might also perform them thoughtlessly and only later bring her actions back to memory and reflect on them. I do not think it will matter to the Manipulation Argument Revisited. The notion of reflective epistemic autonomy is refuted if the monitoring requirement is not met.

What would the basic case look like once we focused on 'free action' as an epistemological problem? The standard debate allows no room for the Monitoring Requirement to perform its job. Here the basic case involves an act that is determined to happen, involves an act that is realised in accordance with the agent's will, but which, unknowing to the agent, was nevertheless determined to happen because some manipulator (for example, Fischer's nefarious neurosurgeon) is making sure that this action will take place, even if the agent would have changed his mind. And so, the monitoring requirement cannot be tested in the standard debate because its basic case is already committed to the assumption, next to other assumptions, of course, that the agent's will is effective. The monitoring requirement adds nothing here.

For the Monitoring Requirement, we need a different type of basic case. If it holds, people should be able to tell the difference between situations in which they do act and those in which they do not act. The basic case that could raise doubts about the success of this requirement is one in which people seemingly perform an action, while in fact, they do not, but nevertheless claim to have performed an action. Psychologists agree that in abnormal cases, as in the case of the alien hand syndrome, people may sometimes perform involuntary movements that, to people themselves, seem like purposeful action.⁷⁶ But psychology also provides several experiments that illustrate normal people may also suffer from a failure of the monitoring requirement. I will describe just one experiment that has been performed by Wegner and Wheatley (1999).⁷⁷ The task participants were asked to perform was to move, together with another participant, a cursor around a computer screen that displayed several objects (a hat, an apple, and so on). Meantime they heard over the headphones a constant stream of words. The fellow participants were really an accomplice in the experiment: they were in complete control and moved

76 This syndrome is discussed by King and Carruthers (2012:204-205).

77 For more examples of experiments that involve technical manipulation of both sensory and emotional experiences: see Metzinger (2009) and Doris (2015).

the cursor to preselected objects at certain times. When the participants saw the cursor move towards an object without first hearing the name of that object, they did not experience control. But when they saw the cursor move towards an object (a movement which, of course, was subtly manipulated by the accomplice) and then heard the name of the object, they nevertheless reported they had experienced control over the cursor. Apparently, the participants were not able to tell the difference between a purposeful movement and a mere movement. Even more so, in these and other experiments, people tend to confabulate explanations.⁷⁸ This experiment has more often been mentioned to argue for people's failure to introspect, for the illusionary character of conscious control over actions, more specifically: for the failure of the monitoring requirement.⁷⁹ It is the social-psychological variety of the philosophical manipulation argument.

Experimental results may show that people are sometimes mistaken when they reflect on the perceptual evidence for their agency. But it does not necessarily show that people are mistaken all the time. It may still be maintained that '[t]he mere fact that error sometimes occurs is not, by itself, reason to suspect that error occurs in every case involving the feeling of agency in reasoning' (Kornblith 2012:152).⁸⁰ To understand how the cursor experiment can be interpreted as *showing* the failure of the monitoring requirement, I will discuss a cluster of theories that underpins much of these experiments and which has helped to explain the incongruities between what subjects in experiments report about their behaviour and what they actually do (or fail to do). They belong to the most important developments in social psychology. These theories are known under the umbrella term 'dual-process theory', although the term is also used for some specific theories.⁸¹ Their core idea is that people's mental processes, the formation of, for example, either beliefs or their will, can be realised in two different ways, which are sometimes called 'System 1' and 'System 2'.⁸² A series of contrasting terms are commonly used to characterise the differences between System 1 and System 2. They are described as respectively: automatic vs controlled, fast vs slow, parallel vs serial, associative vs rule-governed, emotional vs emotionally neutral, effortless vs effortful, inflexible vs flexible, undemanding of working memory vs demanding of working memory, implicit versus

78 'Confabulation' is a technical term in psychology for misrepresentations of memory, usually where people suffer from problems with their memory due, for example, to brain injury, dementia or psychosis. The term is also used more generally – and this provides the relevant context for my discussion – for healthy people, in which case confabulation refers to (incorrect) reasons that people make up in trying to explain their own behaviour when actually it had other causes. Relevant for this context is, e.g. Nisbett and Wilson (1977). For a discussion of both clinical cases and cases of healthy people: see, e.g. Doris (2015:81-99). Also, see Gopnik (1993).

79 See e.g. Bargh and Ferguson (2000:940); Carruthers (2009b, 2010); Caruso (2012).

80 For a similar remark: see Carruthers (2010).

81 For overviews of dual-process theory: see, e.g. Evans (2008); Frankish and Evans (2009); Gawronski and Creighton (2013); Gawronski, Sherman and Trope (2014).

82 This particular terminology was forged by Stanovich (1999). Whether we can really distinguish two systems (two processes, two attitudes) is still an ongoing discussion: see, e.g. Moors and De Houwer (2007).

explicit and finally as unconscious vs conscious.⁸³

For my present argument, we need to know how the terminological pair unconscious/ conscious functions within dual-process theory. What is its status compared to other terminological pairs? Is it the case, for example, that this distinction can be used only sometimes to describe differences between both systems, depending on the type of experiment or the research domain of this experiment, whereas other distinctions describe consistent differences? These questions are part of more general discussions on refining the framework of dual-process theory: they turn on how all these distinctions can be mapped onto each other. Do they coincide? Do they correlate? They have implications for how research is performed and described.⁸⁴

The general difference between both systems is often described in terms of the automatic/ controlled distinction: 'A central characteristic of dual-process theories is that they are concerned with the question of whether the mental processes underlying social behavior operate in an automatic or nonautomatic fashion' (Gawronski, Sherman and Trope 2014:5).⁸⁵ But the literature on dual-process theory shows a tendency to value the unconscious/ conscious distinction as more basic.⁸⁶ This has not always been the case. During the second half of the twentieth century, particularly experimental psychologists gradually turned away from behaviourism. They were again looking for a model of the mind.⁸⁷ As it became clear from experiments that people were not always conscious of what was going on in their perception of their behaviour, psychologists needed a new terminology to express their findings. Freud had, of course, contributed to making 'the unconscious' a new notion in psychology. But among these experimental psychologists, there were few who 'wanted to risk the accusation that they were, God forbid, Freudians' (Wilson 2002:4). The dissatisfaction with Freudian theory probably prevented the term 'unconscious' from immediately becoming currency in dual-process theory.⁸⁸ At first, other notions were therefore introduced, which even varied among the different research domains. The use of these various terminological pairs is partly due to different research traditions from which dual-process theory developed. One tradition that focussed on selective attention and short-term memory was concerned with descriptions that turned on the 'inescapability' and 'efficiency' of, in this case, selective attention and short-term memory. And therefore, the relevant descriptions were automatic vs controlled and efficient vs non-efficient. Drawing on a different tradition, however, one concerned with implicit memory, 'in the following years, the automatic/controlled distinction

83 For even more contrasting terms, see, e.g. table 1.1 in Frankish and Evans (2009:15) and Carruthers (2009:109).

84 As Hofmann and Wilson 2010:198 observe: 'Whereas these features may often coincide, an all-or-none view of perfectly correlated features is clearly an oversimplification'.

85 Also see Bargh (1994); Moors and De Houwer (2006).

86 See Hassin, Uleman and Bargh (2005); Dijksterhuis and Nordgren (2006); Bargh (2007).

87 This first phase after rejecting behaviourism has also been called 'The New Look 1': see Bruner (1992).

88 Hassin, Uleman and Bargh (2005).

often gave way to the dichotomy between explicit and implicit processes, interpreted as synonyms for the terms conscious and unconscious' (Payne and Gawronski 2010:4). To pinpoint more exactly what is supposed to be an essential difference between System 1 and System 2, psychologists also employed the distinctions aware/unaware, intentional/unintentional.⁸⁹ In a series of articles, Kihlstrom (1984, 1987, 1990), and after him, Greenwald (1992), paved the way for reintroducing the distinction conscious/unconscious and use the term 'unconscious' to describe a phenomenon of the mind that is by now recognised as obviously different from the Freudian unconscious: the 'cognitive unconscious' (Kihlstrom 1987), 'psychological unconscious' (Kihlstrom 1990), 'adaptive unconscious' (Wilson 2002), 'new unconscious' (Hassin, Uleman and Bargh 2005). Terminological issues are still a matter of debate, including what the 'unconscious' means as opposed to other qualifications such as 'automatic'.⁹⁰ At the same time, the choice for any pair of contrasting terms to describe the general difference between System 1 and System 2 is also 'a matter of convention' (Payne and Gawronski 2010:5). At least when psychologists turn their attention to philosophical issues on consciousness, they tend to give priority to the conscious/unconscious terminology over other terms, including the automatic/ controlled terminology. This clearly serves polemical purposes, namely to underline a psychological view of the mind that resists philosophical views of the conscious as the seat of control, freedom, rationality.⁹¹ It also holds for the debate on responsibility for implicit bias. Even when in descriptions of implicit biases, the adjective 'implicit' has become the standard rather than 'unconscious', several aspects of implicit bias can be understood in senses of the unconscious that are relevant to the debate on responsibility.⁹² But whatever terms are employed in describing differences

89 Using the terms 'intentional-unintentional' as running parallel to aware-unaware or conscious/unconscious has been unfortunate from the perspective of philosophical discussions, which recognise the ambiguity of these terms because of the difference between their ordinary and philosophical use. Schmid and Schweikard (2009:14-15) observe that this ambiguity seems to be typical for the English language (in contrast to, for example, the German language). In its philosophical use, both conscious and unconscious mental states can be intentional. Intentionality is, therefore, in this sense, no distinctive feature. At first, social psychologists did not seem to recognise this ambiguity. See, e.g. Lewis (1990). Bargh takes over the intentional/unintentional terminology, while Chapman (1990) already warns for its ambiguity. Social psychologists have gradually appropriated the relevant philosophical literature. An anthology devoted to automaticity, published in 1989, still carries the title 'Unintended Thought'. Here the term 'unintended' (sometimes 'unintentional') is used as a synonym for 'unconscious' and in contrast to 'intentional' (sometimes 'intended'). But the title of its successor has changed into 'The New Unconscious' (2005). Also e.g. Malle (2005); Moors and De Houwer (2006).

90 For a general discussion: Bargh (1994); Moors and De Houwer (2006). For a critical discussion of equating unconscious with subliminal: Bargh and Morsella (2008). For a discussion of the term 'unconscious' as referring to different aspects of attitudes and psychological processes: Gawronski, Hofmann and Wilbur (2006); Hofmann and Wilson (2010).

91 E.g. Higgins and Bargh (1992); Bargh and Morsella (2008); Hofmann and Wilson (2010).

92 Important for introducing the terms of implicit/explicit to describe social cognition: Greenwald and Banaji (1995). Banaji has been involved in much of the research on implicit bias. For a discussion of the unconscious aspects of implicit attitudes: Gawronski, Hofmann and Wilbur (2006). For a discussion of implicit bias in terms of automaticity: Devine (1989).

between System 1 and System 2, they are relevant to the debate on responsibility and agency to the extent that they imply the principal inaccessibility of (certain information, features, et cetera, of) System 1 to System 2 or, to put it differently, of certain beliefs, volitions and other attitudes to people's consciousness.

Dual-process theory does not necessarily imply the failure of the Monitoring Requirement. To understand how it may function as part of a deterministic view on the mind, we need to inquire into the exact relation between System 1 and System 2. The two systems usually seem to be acknowledged by social psychologists as simply two different types (processes, attitudes), each having its own features, its own weaknesses and strengths. Although it is accepted that the unconscious mind bypasses the conscious mind (sometimes, often), some psychologists regard both systems as 'parallel competing processes' (Evans 2008). But this idea raises serious doubts:

'It is generally assumed that System 1 and System 2 are (largely) distinct from one another. But then one immediate challenge for dual systems theory concerns the relationships between the two (sets of) systems. How, if at all, do they interact with one another? How is it possible for System 2 to override the operations of System 1 in controlling behavior? And how are we to imagine that System 2 could have evolved? If System 2 is charged with generating new beliefs, desires, and decisions, then how could it have evolved alongside a set of mechanisms (System 1) that already possessed precisely those functions?' (Carruthers 2009:111)

Below I will sketch an alternative view on the relation between System 1 and System 2, for which I will draw on Carruthers, one of the few philosophers that have explored dual-process theory from a philosophical perspective.⁹³ Crucial for our discussion is that it includes an argument against higher-order accounts of consciousness and therefore provides a theoretical basis to interpret the cursor experiment as symptomatic for the failure of people's introspection. Carruthers targets one feature that usually characterises higher-order theories:

'most higher-order accounts of consciousness can be described as warranting a belief in *introspective* access to our own experiences (and by extension to our thoughts), in the sense that the relationship is held to be especially immediate and direct. On this view, the access that we have to our own experiences (and thoughts) is radically different from the sort of interpretative access that we have to the experiences and thoughts of other people' (King and Carruthers 2012:208).

93 Some philosophers do make distinctions that resemble those between System 1 and System 2 (see Frankish and Evans 2009:19-21). Carruthers sometimes seems to present his argument as one against dual-process theory. As I will argue, his problem is not with System 2 as such (Carruthers does not deny conscious thinking), but with an interpretation of System 2 as somehow in control.

Interestingly he mentions amongst others both Frankfurt's (1971) and Moran's account (2001) as exemplary for such views. Against them, Carruthers denies that people have this immediate and direct access to their mental states.⁹⁴

To explain Carruthers's argument, I will begin by introducing two distinctions that are important to it. First, he distinguishes between (quasi-)perceptual events and propositional attitude events. To the first category belong 'the subject's current circumstances, or the subject's current or recent behavior, as well as any other information about the subject's current or recent mental life' (2009b:3). These include, for example, perceptual and mental imagery (visual, but also auditory imagery that is used in 'inner speech'), proprioceptive data, emotional feelings and bodily sensations. To the second belong believing and willing (i.e. the cognitive and conative mental states). He also distinguishes between introspection and attribution. The two notions imply a different conception of how people come to have knowledge of (perceptual or attitude) events. The first implies immediate access, while the second implies mediated access to events, i.e. these events are not directly 'perceived' but only attributed.

We now can state some of the claims he makes. Carruthers claims that people have introspective access to (quasi-)perceptual events. This causes the formation of perceptual beliefs. Another claim is that people do not have meta-cognition: they do not have introspective access to propositional attitudes, such as willing. The (quasi-)perceptual events are all that is available to them. This claim implies that people can see, for example, their arm raising but not 'see' the decision that prompted raising this arm. But people do have beliefs about what raising this arm means. Carruthers's claim that people, as they have not immediate access to propositional attitudes, need to take recourse to an interpretation of what they perceive (the (quasi-)perceptual events) in much the same way as they would interpret other people's behaviour. Taken together, these claims would predict that people in a situation where their finger or hand is being moved by an invisible external cause, at least not by a decision of their body, would still interpret this as a movement of their own and would attribute to themselves a decision. This self-attribution would cause them to believe that they have decided to move their finger or hand. Carruthers would therefore also claim, I think, that the monitoring requirement is not met because people's process of belief formation about their own physical acts cannot, in principle, be a reflective warrant for monitoring these acts.

Carruthers sustains his claims with a view of the mind's architecture that aims at improving the theoretical framework for experimental psychological research. He acknowledges that its architecture can be made sense of in terms of the distinction between System 1 and System 2 processes. He also agrees that their differences can be

94 He criticises them in resp. (2007) and (King and Carruthers 2012). Kornblith, another philosopher interested in dual-process theory as supporting claims against introspection, is critical of Frankfurt and Moran: Kornblith (2012, resp. sections 3.1-2 and 3.3). Pereboom takes Frankfurt's account (1971) as exemplary. In none of his articles, to my knowledge at least, does Pereboom address or even mention Moran's account (2001).

described by contrasting them in terms of fast-slow, parallel-serial, and finally also in terms of unconscious-conscious. But he disagrees with some social psychologists about how to conceive the relationship between both systems. He denies, for example, that System 1 and System 2 represent two independent mechanisms.⁹⁵ Carruthers claims that System 2 is realised ‘in cycles of operation of System 1 (rather than existing alongside of the latter)’ (2009a:112). To understand what this claim means and why it undermines any claim of introspecting the propositional attitudes, we should get a closer look at what is inside System 1. This is not one single system, but a set of systems. Among them are a range of perceptual systems (visual, somatosensory, auditory) and a range of conceptual systems. The perceptual systems ‘broadcast their outputs’ to the conceptual systems, some of which in their turn generate beliefs, ‘some create new goals, and some generate decisions and intentions for action’ (2010:79). When these beliefs, decisions and intentions (presentations) have been generated, they can be stored in memory and again be activated during reasoning. The conceptual systems include more than one belief-generating system. For example, one system generates, based on outputs from the perceptual systems, first-order beliefs. Another one is the mind-reading system which can generate higher-order beliefs.

The mind-reading system is constitutive of people’s consciousness. When people reflect on either their perceptions or their own mental states, the underlying processes take place within this system. Carruthers denies that the outputs of the systems that generate first-order beliefs and decisions have causal pathways towards the mind-reading system. This agrees with his claim that people have no access to propositional attitudes, including first-order beliefs.⁹⁶ The mind-reading system does have access, however, to outputs from the perceptual systems. When it comes to perceptions that do not necessarily appeal to people’s own behaviour or circumstances, as in this example, it is not immediately relevant why we should distinguish between two belief-generating systems that both operate based on perceptual events. When, for example, people perceive a dog chasing a ball (or, in Carruthers’s technical vocabulary: when they receive ‘as input a visual representation of a dog chasing a ball’), in both systems, the belief is generated that ‘I am seeing a dog chasing a ball’.⁹⁷ This would qualify as introspection. As such moments we are conscious of what we see: a dog chasing a ball. A difference between both systems arises once people during belief formation need to make an appeal to their own behavioural circumstances. To understand what they perceive, they are then required to reflect on their own mental states (either first-order beliefs or decisions). The reflection on one’s mental states takes place within the mind-reading system. As we have already remarked, Carruthers denies that the mind-reading system has access to systems that generate first-order beliefs and decisions. But the mind-reading system provides

95 He ascribes this view to Nichols and Stich (2003).

96 Carruthers (2010:81).

97 For more explanation on how the mind-reading system is able to conceptualise perceptions: Carruthers (2010:80-81).

a solution to this problem because it underlies people's capacity to attribute mental states. Interestingly an evolutionary link is suggested between mind-reading and manipulation as a social ability (see King and Carruthers 2012:212, note 25) because as cooperation between people became more complex, they had to improve their ability to 'reading' other people's minds, i.e. of attributing beliefs and decisions to other people. In this way, the mind-reading faculty enabled people 'to anticipate and manipulate the behavior of other people, as well as to cooperate more effectively with them' (King and Carruthers 2012:212). As people try to grasp their own behaviour consciously, they need to rely on the same mind-reading powers as they use for reading other people's minds. This process qualifies as self-attribution. As the beliefs it produces are not about what people perceive but about people's mental states, they count as higher-order beliefs (meta-representations).

Much of the processing that is going on in the various systems discussed so far should be understood as unconscious as these systems are still located within the System 1 domain. System 2 is put to work when people reflect on either their perceptions or their own mental states. But the space allowed for reflection is narrow. As I discussed earlier (chapter three), consciousness can be understood in two different senses: transitive and intransitive. Reflection would be the process of a transitive consciousness: an unconscious mental state becomes conscious when 'it is taken as the object by a higher-order state' (Zahavi 2007:268). In Carruthers's account of the mind-reading system, issuing in conscious mental states, we can see an oscillation between both senses of consciousness. When people reflect on their own beliefs and decisions, they appear to be conscious in a transitive sense. But rather than accessing their own mental states, their reflection is a case of swift self-attribution, which gives it the sense of introspection.⁹⁸ Even reflection as self-attribution does not take place as an autonomous conscious process because much of the mind-reading system upon which it relies operates unconsciously.⁹⁹ This is illustrated in an example Kornblith provides of a woman called Mary who has misplaced her keys (2012:144):

'In getting ready to leave her house, she looks in her purse, where she usually leaves them, and discovers that they are not there. She then engages in a process of self-conscious reasoning, that is, the kind of reasoning which goes on in System 2. When did I last use my car keys, she asks herself, and what did I do with the keys when I last arrived home? As she consciously raises these questions, she mentally rehearses the path she followed from her car to the house, focusing her attention, at each point, on what she might have done with the keys.'

98 Admittedly Carruthers does not employ himself talk of transitive/intransitive consciousness but follows Block (1995:232-233) in distinguishing between phenomenal consciousness and access consciousness. For more comments on these distinctions: see Macdonald (2008:321-322, note 8). I do not think that it matters for my discussion (cf. Zahavi 2007, section 1).

99 For a more detailed explanation: see Carruthers (2009a:120).

Her reflection (which consists in conscious reasoning) allows her to analyse several options of where she could have left her keys. Although we may understand this reflection as a process that Mary controls, it is clear, as Kornblith argues, ‘that the conscious contents of her mind cannot exhaust what goes on in her reasoning’ (idem). In other words: she cannot control the unconscious processes that nevertheless help her to rule out some unlikely options: ‘Mary does not self-consciously entertain the thought that her keys cannot travel discontinuously through space and time’ (idem). Of course, she believes this, but this belief does play a conscious role in her reflection. The case of Mary’s reflection only shows what is common about people’s reflections which is that they rely on a background of beliefs that are unconscious and that are accessed unconsciously: ‘when a large body of background beliefs play a role in determining what an agent believes, System 1 is inevitably involved. Reasoning in System 2 is thus influenced by reasoning in System 1’ (Kornblith 2012:146). Carruthers phrases it even more restrictive: ‘System 2 is realized in cycles of operation of System 1’ (Carruthers 2009a:124).

The previous explanation still leaves room for introspection. Carruthers and his co-author allow for the possibility that people do have genuine introspective access towards their attitudes:

‘From the fact that we sometimes engage in swift unconscious self-interpretation it doesn’t follow that we always do, of course. On other occasions we might have access to our attitudes that is introspective in character. It is important to see, however, that from the subject’s own perspective there is no epistemic difference between the two. Since it can seem to subjects that they are introspecting their attitudes when they are demonstrably (but unconsciously) self-interpreting, the common-sense intuition that we have immediate introspective access to our own propositional attitude events of deciding, judging, and so forth should arguably be given little weight’ (King and Carruthers 2012:214).

Carruthers’s decisive argument is, therefore, that people are not able to tell the difference between introspection and self-attribution (or: self-interpretation). If people engage in self-attribution, it happens so fast that it ‘feels’ like genuine introspection. And so, we must conclude, the Monitoring Requirement is not met. Responsibility is impossible.

4.6. Conclusion

I have argued that the liberal-mentalist notion is not able to deal with agential self-ignorance. One strategy to show this was to analyse whether people can detect their own

implicit biases. While this provides a *local* problem for assessing responsibility, exploring the underlying mentalist assumptions led me to pursue a different strategy and analyse what poses a more radical challenge: the *global* problem of agential self-ignorance, better known as the problem of determinism. It arises once liberal theorists accept an ontological (or, as I explored it here, *neurological*) interpretation of the conscious-unconscious distinction. In the standard debate on determinism, thought experiments centring on manipulation provide an important argument against the claim that people have freedom and, therefore, responsibility: the Manipulation Argument. I discussed two versions. The first version focuses on the *volitional* conditions for action: what kind of control over one's will guarantees that action is truly free? The second hones in on the *cognitive* conditions for free will: what kind of control over one's beliefs guarantees that a will is truly free (and consequently, that an action is truly free)? Only the first is extensively discussed in the determinism debate, but as this largely ignores the issue of what it means for people to have introspective self-knowledge and, more broadly, to be knowers, I chose to shift the discussion towards the question of how to understand the idea of introspective control over beliefs. It resulted in a revision of the notion of reflective epistemic autonomy and a second version of the Manipulation Argument. In the end, the two versions of the Manipulation Argument were able to refute the notions of reflective volitional autonomy and reflective epistemic autonomy.

My aim in discussing the problem of determinism has been to show that the mentalist framework on which the liberal-mentalist notion of responsibility relies can be turned against the liberal-mentalist notion of responsibility. But however problematic the liberal-mental notion of responsibility now appears to be, I claim that it is really the mentalist framework itself that is problematic. In the next chapter, I will argue against it by exploring the relation between determinism and people's condition as knowers (or, more traditional terms, as rational beings). At the same time, I will start developing an alternative framework: a pragmatic one based on Robert Brandom's writings.

5

Chapter 5

**A pragmatist framework:
freedom and rationality**

5.1. Introduction

In this chapter, I will shift from the vocabulary of reflectiveness to the vocabulary of rationality.¹ The notion of rationality that still dominates many philosophical discussions is rationality as *deliberation*. It is usually understood, on the theoretical side, as the acquisition of knowledge according to formal reasoning – logic providing the model of formal reasoning and thus of theoretical rationality – and, on the practical side, as acting according to a means-goal or means-desire type of structure: here deliberative acting is acting deliberately.² As practical rationality is understood here only in a specific sense (what could be called the species of the genus ‘practical rationality’), I will reserve for this the term ‘instrumental rationality’. Using now the vocabulary of rationality, we may claim that the condition of people in a deterministic universe is marked by the disunity of rationality (both theoretical and practical) and action because the thinking mind is unconnected to the acting body: the ownership of the actions we think are ours is only an illusion.

The disunity of rationality and action makes no sense, however, as an *absolute condition*. The problem underlying this master assumption is the notion of deliberative rationality itself. Useful as it may be for socio-political analyses, it fails as a basis for the determinists’ discussion on freedom and responsibility. This chapter will argue for a different notion of rationality, which is more basic than deliberative rationality and presupposes the *unity of rationality and action*. For this notion, I will draw heavily on the writings of Robert Brandom, who has worked for many years on the notion of *inferential rationality*. If Brandom’s account is plausible enough, it also provides an argument against the thesis of determinism because denying the unity of rationality and action, which this thesis implies, precludes understanding knowledge, not in the least scientific knowledge. And if we cannot explain the acquisition of knowledge, how could we even know the determinism thesis? Brandom’s account provides, I think, the proper framework not only to overcome the determinism thesis but also for picking up again on the two issues that are impossible to address based on the reflective notion of responsibility (also see the previous chapter, section 1):

- 1 Of course, the notion of rationality was already looming in discussions of the previous chapter because, for example, in the story on Professor Plum, the manipulative intervention in his reasoning process was part of Pereboom’s strategy to also address this type of view, more specifically those of Fischer and Ravizza (1998), and of Wallace (1994). But the perspective of the arguments was still that of reflectiveness. My choice of shifting from *reflective* accounts, in the previous chapter, towards *rational* accounts of freedom and responsibility, in this chapter, will be explained in more detail in section 4.
- 2 In psychological debates on rationality, this has been called ‘the Standard Picture’: ‘to be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability theory and so forth’ (Stein 1996).

1. Are people responsible for their hidden prejudices?
2. Can their hidden prejudices be changed in a morally relevant sense?

In the next sections, I will proceed as follows. First, I will argue that the participants in the determinist debate on freedom who take the reflective will as their starting point – many libertarians, neo-Kantians, soft and hard determinists – use the wrong notion of freedom. They conflate what are really two different levels of analysing freedom, a moral-political and metaphysical level (section 2). Determinists are committed to a theoretical framework that recognises only two types of actions: deterministic (non-responsible) and non-deterministic (responsible). Combining the levels of a two-level notion of freedom allows us to introduce a new category of actions: transcendently free but non-responsible actions (section 3). This will invite us to look for another transcendental notion of freedom. I will introduce Brandom's notion of freedom as rationality first by contrasting it with the standard notions on rationality within the determinist debate (section 4) and then by explaining it as a form of *normative* bindingness (section 5). Central to his account of rationality is a pragmatist notion of normativity. Discussing some criticisms will allow me to clarify that people's status as having freedom consists in their being knowers (section 6). Lastly, as a response to the Manipulation Argument, I will present a counterargument against attempts at naturalising responsibility: the Cognitive Suicide Argument (section 7). It will show that giving up the freedom of agency also implies giving up knowledge.

5.2. Free will as the wrong notion of freedom

Do proponents of the Manipulation Argument succeed in undermining the liberal-mentalistic notion of freedom and responsibility, which has its basis in the reflective will? If so, this is because the liberal notion is primarily a *moral-political* notion, not the type of notion designed to explain freedom in a mechanical world.³ In short: they have chosen

3 In his diagnosis of liberalism as a modern tradition and of its specific notion of freedom/agency as one that understands practical reasoning as basically a volitional one, MacIntyre remarks, although he does not particularly address the determinism debate, that '[c]ontemporary analytic philosophers [...] often take themselves to be representing the timeless form of practical reasoning as such, when they are in fact representing the form of practical reasoning specific to their own liberal individualist culture' (1988:340). From the various authors he mentions as his examples, I take MacIntyre to claim that this deep-seated assumption is part of theorising not just in liberal political theory but also in other philosophical domains within analytic philosophy.

the *wrong* notion of freedom.⁴ To make this claim plausible, I will introduce a two-level understanding of freedom. Next, I will sketch a historical account that indicates how the will came about as the wrong notion of freedom and what ought to be treated as the right kind of freedom. In the next section, I will conclude with a discussion of Frankfurt's account and Pereboom's Manipulation Argument to show the problem with using the wrong notion of freedom.

I suggest that we should understand freedom as a two-level notion, a notion that may function to analyse human action at two distinct levels. At one level, it allows us to qualify actions in terms of the contradictory distinction between 'free' and 'unfree' in the sense of 'undetermined' and 'determined'. At another level, it allows us to appreciate actions in terms of the contrastive distinction between 'free' and 'unfree' in the sense of 'more free' versus 'less free'.⁵ The two levels should be kept apart carefully.⁶ They each indicate a different sort of demarcation between freedom and unfreedom and therefore imply different sorts of notions of freedom (and so of 'unfreedom').⁷ When we talk about freedom at the contrastive level, we employ moral-political notions of freedom. These are usually regulative: they prescribe what we should understand as an impermissible (immoral, illegitimate) constraint on people's ability to act. At the disjunctive level, we use transcendental notions of freedom: they explain what is constitutive of beings that have freedom as opposed to beings that have not. Both levels are conceptually connected: moral-political freedom presupposes transcendental freedom. What matters for the determinist debate on freedom are only disjunctive, i.e. *transcendental*, notions of freedom, notions that presuppose a sharp boundary between what is deterministic, i.e. what can be reductively understood in terms of mechanical causation, and what is not deterministic.⁸

4 For similar criticisms (but from somewhat different theoretical perspectives): see Arendt (2006a); Greshake (1981); Hruschka (1986); MacIntyre (1988); Smiley (1992) and Kobusch (1993). Smiley, for example, claims that 'the modern concept of moral responsibility is not, like its Classical and Christian counterparts, internally coherent' (1992:73) because it 'enables us to conflate our practical judgments of causation and blameworthiness into one ostensibly factual discovery of moral responsibility' (1992:167). I, too, am critical of this 'modern concept', but I follow a different strategy. I focus on the notion of causality itself, i.e. a notion that analyses people's practices in terms of a (presumed) link between a will and an action, and claim that this notion already presupposes a more fundamental, moral-political notion, as I will explain in this and the next sections.

5 A familiar example of a two-level notion is 'moral', which can be understood in a contrastive relation with 'immoral' and in a contradictory relation with 'non-moral' or 'amoral'. I do not know of any technical name for these sorts of notions, so it seemed convenient to me to invent one.

6 Schnädelbach (1992b) analyses the predicate 'rational' in a similar way. Also, see section 5.

7 A conceptual confusion may invade our use of two-level notions if it is not clear in relation to which other terms they are used: 'moral' in relation to 'immoral' or to 'non-moral'? To be clear about the level at which I use a notion of freedom, I will use the verb 'have' to indicate use at the contradictory level – therefore: having freedom – and the verb 'be' to indicate use at the contrastive level – therefore: being free.

8 Perhaps the choice for the adjective 'transcendental' is an unfortunate one. But I hope that the argument in this section makes it clear enough how I intend to use 'transcendental freedom'.

Addressing the issue of freedom at either level of analysis raises at least two questions. First: what is the ‘thing’ to which freedom is either ascribed or denied? The answer to this question is the same for both levels of analysing freedom if we treat them not as simply two different levels but as being connected by some ‘hinge’. Once we have settled what the ‘hinge’ consists in, we arrive at a second question: how is it possible that ‘it’ (whatever this hinge is) is (more) free (or: unfree, less free)?

In the previous chapter, Frankfurt’s notion of freedom was discussed as exemplary for a family of notions of freedom that determinists often take as the starting point for their debates. These notions have in common that they provide the same answers to the two questions just mentioned. The hinge between both levels, according to these notions, is the will. What enables the will to be free is for each of these notions a reflective process of some kind, even if these notions sometimes pick out different criteria that turn this reflective process into one that ‘liberates’ the will. The criterion Frankfurt mentions, for example, is identification, a feature that is typical for notions of personal autonomy. Whatever variations distinguish these notions, their common core in defining freedom is, therefore, the ‘reflective will’. Determinists have chosen the notion of *freedom as reflective will* for debating the challenge of determinism. The problem is that they have chosen the wrong notion.⁹ It conflates the two levels of analysing freedom because determinists treat as a transcendental notion what is really a moral-political notion.¹⁰ The result is an inconsistent notion that is unfit to face the challenge of determinism. The conflation of levels is best explained by turning to a brief historical account, for which I will use the two-level notion of freedom as a piece of meta-language.

The current determinist debate on free will is presented mostly as a continuation of older debates in the early modern period, the Middle ages and even Antiquity.¹¹ Modern determinists sometimes try to classify theorists from those earlier periods as either ‘soft determinists’ (‘compatibilists’) or ‘hard determinists’.¹² Even historians of philosophy have taken over this vocabulary to make similar attempts.¹³ Such attempts presuppose that the continuities between modern and earlier debates on determinism are more relevant than their discontinuities. I will mention three presumed continuities:

9 Notions of personal autonomy, which have been inspired by amongst others Frankfurt’s account, qualify as belonging to the moral-political level, as I have already argued in chapter three.

10 I think that the ambiguity of Frankfurt’s notion of freedom is apparent from the influence it had in very different debates (see chapter 2, section 6). Other notions of freedom too met the fate of being shifted from one level to another, as in the case of T.H. Green’s notion of freedom as ‘self-determination’, which was intended as a moral-political notion, but was nevertheless used by William James (1884) as an example of a transcendental notion. And it might be argued that T.H. Green’s moral-political freedom was itself the result of treating Hegel’s text on the master and his servant as a moral-political analysis (Wempe 2004) what is really a transcendental analysis (Ottmann 1981). In current debates on determinism, T.H. Green’s notion does not play a role anymore.

11 See, e.g. Pereboom (2009); Nichols (2007).

12 Pereboom (2009).

13 Kenny (1973).

1. The problem of determinism is one about the demarcation question. If people are transcendently free – in the sense of *not being determined* (whatever this may mean) –, this explains their exceptional status compared to the rest of the natural order. Or perhaps it should be reversed: if it can be explained what their exceptional status consists in, then this is what we call transcendental freedom.
2. Freedom [of action] precedes morality [being a moral subject], sometimes expressed in the claim that free will precedes moral responsibility.¹⁴ The reason for this, as we have seen earlier, is that the fate of morality is tied up with the fate of the demarcation question. If it cannot be explained that human beings have an exceptional status (i.e. are free), then moral responsibility is impossible.
3. What functions as a ‘hinge’ is the will.

I take the opposite view: discontinuities in the history of debates on determinism are more important.¹⁵ Ignoring this is bound to distort our understanding of early modern philosophers (and therefore also of modern philosophers who came later, such as Kant and Reid) and their debates on determinism. And this, in its turn, will prevent us from understanding what assumptions underpinned their use of the ‘will’ as a philosophical notion and how these were lost sight of as early modern debates transformed into our present debates on determinism. I would like to suggest a radically different reading of the early modern debate on determinism – or, what was then the common term, *necessity* (and its theological equivalent: predestination) – by claiming that what previously were called continuities are really discontinuities.¹⁶

1. The first discontinuity is a shift from the question *of how* transcendental freedom works towards the question *of whether* human beings do have transcendental freedom. While the demarcation question is the primary concern for modern determinists, it was *not* for early modern necessitarians.
2. The second one is a shift from taking morality as preceding [in the sense of constituting] [individual] transcendental freedom towards taking transcendental freedom as preceding morality.
3. The key to understanding both discontinuities is a third discontinuity that is more fundamental, which is the shift from ‘person’ to ‘will’ as what necessitarians/determinists recognise as the hinge between both levels of freedom.

¹⁴ Also, see chapter 4, section 2.

¹⁵ Smiley (1992) shares this view and also criticises the current use of the notion of free will in debates on determinism. She chooses, however, to contrast the modern notion of responsibility with Classical and Christian notions, whereas I try to locate a discontinuity between the modern notion and the early modern *contractarian* of responsibility.

¹⁶ It seems to me that this focus on discontinuities is helpful for a reconstruction of the *history* of debates on free will. In this respect, Pereboom’s introduction (2009) to an anthology of historical key texts for these debates seems oddly *ahistorical*.

Notions of personhood are usually traced to the ancient tradition of ‘persona’, a term that is traditionally thought to be the Roman translation of the Greek *πρόσωπον* (face). Due to its metaphorical use in the theatre, it produced the meaning of ‘person’ as ‘mask’.¹⁷ Later this developed into the related but distinct meaning of ‘role’. Its use was not limited to a dramaturgical setting but acquired, for example, within the Roman legal context the more specific and technical meaning of ‘legal subject’.¹⁸ To these meanings of ‘person’, the scholastic tradition added two different meanings: ‘person’ as ‘substance’ and as ‘moral entity’ (*ens morale*).¹⁹ Being a moral entity was intimately linked with being rational to the extent that these two dimensions were only different sides of the same coin: being rational, deliberating, was having access to knowledge about moral actions.

The interest of moral-political theorists in the early modern period was mostly with analysing actions rather than with analysing the notion of personhood as the source of these actions.²⁰ But they nevertheless needed a notion to support their analyses of actions. This was the notion of personhood as they had inherited it from the scholastic tradition. The scholastic notion of personhood as ‘substance’ and moral entity originates in discussions on the nature of the trinity. These took as their starting point the question of how we should understand Jesus as having both a human and divine nature.²¹ This way of thinking was then applied to human beings: they were understood not just as natural beings but also as persons, i.e., as rational and moral beings. Boethius provided the influential definition that links the person as being a substance to rationality: ‘Persona est naturae rationabilis individua substantia’. But being rational was not yet understood as being free in any metaphysical or transcendental sense. In the thirteenth century, discussions on the trinity gave rise to the meaning of person as ‘ens morale’, which for the first time implies that people, as persons, belong to a different domain of being: the *esse morale*.²² It implies a break with the Aristotelian metaphysics that conceives people as belonging to the same domain of being as ‘things’ and which also still underlies Boethius’ definition of personhood. In a way, this introduced freedom as a two-level notion: freedom as opposed to the domain of nature’s necessity (transcendental freedom) next to freedom within the domain of morality (moral-political freedom). The notion of ‘persona’ functioned to keep these two levels apart and to secure transcendental freedom. And here, surprisingly from a modern perspective, morality, in the sense of having moral status, was understood as preceding freedom (of action).

Sometimes a distinction is made between efficient causes and final causes and

17 Cotta (1983:160).

18 E.g. Stagl (2012). Cicero, too, used the meaning of person as ‘role’ in a metaphorical sense: see Gill (1988).

19 Kobusch 1993 and Thiel 2011 (esp. the introduction).

20 See chapter 2.

21 So the ‘person’ as a being that is responsible for actions seems to have two different conceptual origins: a legal and a theological one.

22 Kobusch (1993).

consequently between efficient necessity and final necessity.²³ This is only a very abstract distinction. Efficient necessity can, however, be conceptualised in more specific ways, for example, as either fatalism (in Antiquity) or mechanical determinism (in present debates). Whereas efficient necessity is almost the only type of necessity that matters to today's determinists, for medieval (and very early modern) theorists, it was almost non-existent.²⁴ For them, the demarcation between human beings and animals was therefore not one between free and (mechanically) determined, but between having a status as a moral-rational entity and not having that status.²⁵ Instead, they were concerned with final necessity. This followed from their own characterisation of people as moral-rational beings. Whereas rationality and moral responsibility were assumed to form an unproblematic unity, from the late thirteenth century onwards, some theorists began to claim that rationality conflicts with moral (personal) responsibility. If rationality provides the source for moral reasons, people have a moral reason to do *X* and therefore have a compelling motive to do *X*, but then they have no real personal responsibility.²⁶ People's rational nature (in the sense that medieval theorists understood it) therefore amounts to rational determinism.²⁷ The consequence of this view, however, was that it became problematic to understand how people who were supposed to have moral knowledge (and therefore to act in agreement with this knowledge) could still act contrary to their moral knowledge. This is the problem of weakness of will (*ἀκρασία*).²⁸ The notion of 'will' as a distinct faculty was introduced to solve precisely this problem.²⁹ (Only later, as we now know, the will as a distinct faculty became itself a problematic source of

-
- 23 Cf. Normore (2002), who uses these phrases: 'final causal determinism' versus 'efficient causal determinism'. In a note, Sir William Hamilton added to a letter by Thomas Reid (1793/ 1872:87, note †), he remarks: 'There are two schemes of Necessity – the Necessitation by efficient – the Necessitation by final causes. The former is brute or blind Fate; the latter rational Determinism. Though their practical results be the same, they ought to be carefully distinguished in theory.'
- 24 Adamson (2014).
- 25 This is a considerable difference with modern determinist debates: here, the key term to define the possible demarcation between human beings and mechanical objects is freedom rather than rationality and/or morality.
- 26 Cf. Kent (1995, chapter 3) and Hoffmann (2014).
- 27 Normore (2002; also see 2007). Rational determinism seems to be no subject of discussion in modern debates. For an exception: Bourke (1938). Rational determinism, in this late medieval version, is also an ethical determinism (a view of this name is ascribed by Taylor (2006) to Plato) because it is concerned with the reasons people for doing the right thing. In Aristotelian terms, the medieval notion of determinism was about final causes rather than efficient causes. For medieval theorists, determinism in terms of efficient causes was not a plausible option (Adamson 2014).
- 28 Kent (1995). Today the problem of weakness of will is often defined in a utilitarian way: acting against one's own interest (in contrast to older definitions: acting against one's moral knowledge).
- 29 The notion of will (*voluntas*) had played virtually no role in medieval discussions until the thirteenth century. Afterwards, a shift seems to take place from using the term 'liberum arbitrium' towards using the term 'libertas voluntatis'. Both terms are usually translated into English as 'free will'. Kent, however, draws attention to the fact that this translation is problematic for 'liberum arbitrium' (1995:98–99). It is therefore equally problematic to claim that philosophers using this term were debating 'free will'. This shift in terminology presents yet another discontinuity in the history of discussions on 'free will'.

transcendental freedom.) In the various accounts that were offered for understanding the relation between reason and will, the ‘will’ did not function, however, to demarcate human from non-human beings: ‘the fundamental issue was not whether human beings are free, but whether intellect or will is ultimately responsible for their freedom’.³⁰ Once again, for so-called voluntarists, the issue of *whether* people are free was settled by their status of being a person, i.e. an entity having moral rationality, while the notion of ‘will’ functioned to secure moral responsibility (in the sense of *personal* responsibility) against the necessitation by rationality.

The notion of a person according to late medieval insights – the person as a substance, as a moral-rational being (*ens morale*) and as having a will (the guarantee for the responsibility she can call her own) – still played an important role in the transition towards the early modern period, even though it was more often assumed than analysed or discussed. Some early modern theorists, however, had their own reasons to adopt this notion, specifically to reject the characterisation of personhood as substance, which was a metaphysical status.³¹ The notion of personhood was developed in two different directions that both moved away from ‘person’ as substance: personhood as self-consciousness and personhood as a result of imputation.³² The first made it possible to develop notions of self-governance, as we have seen in chapter two, but also to theorise ‘individuation’ and ‘identity’.³³ The second proved useful, especially for early modern contractarian theorists.³⁴ The imputation model was realised in part by reintroducing Roman-legal thought, which offered a notion of personhood that was social rather than metaphysical or ontological.³⁵ Hobbes is perhaps the first to have introduced the notion of ‘person’ in terms of imputation (or: ascription).³⁶ After him, Pufendorf explored more explicitly and more fully the notion of ascription, which in the long turn proved to

30 Hoffmann (2014:414).

31 For example, Pufendorf tried to detheologise juridical and political theory. See Hunter (2001).

32 See, e.g. McKeon (1957).

33 Thiel (1997 and 2011:18).

34 See McKeon (1957).

35 See Thomas (1998); Pettit (2008); Thiel (2011). Although it is possible to understand personhood, as it is understood in the tradition of self-consciousness initiated by Locke, in terms of the tradition of imputation and therefore as a form of self-imputation (Herbert 1996), this would therefore obscure a crucial difference between both traditions. What a person is, according to the second tradition, depends on a community. For the tradition of self-consciousness, it is the person itself.

36 Pettit (2008); Milanese (2014). As part of the *continuity* view on the history of determinism debates, Hobbes is usually presented as a compatibilist because he believes both that human beings are determined, and that morality (responsibility) is nevertheless responsible. But this ignores various dimensions of his views on man. First, his materialistic view can also be considered as a choice that was convenient for certain political arguments (Jackson 2007). Second, Hobbes claimed that people, as persons, differed from animals in having access to language, which he considered to be a human invention (Pettit 2008). Perhaps he owes this view to the scholastic tradition even though he used the Roman-legal tradition to support his notion of personhood. Thirdly, his determinism has also been interpreted as a logical necessity, which would bring Hobbes closer again to the scholastic tradition.

be influential in criminal law.³⁷ At the same time, the early modern use of this notion preserved the scholastic meaning of moral entity as one that marked the boundary between human and non-human beings.³⁸ In this respect, the early modern notion of personhood differed decisively from the Roman-legal notion of personhood.³⁹ As I previously explained (chapter 2, section 4), the imputation strategy offered obvious advantages as it provided a more complex vocabulary for ascribing and assessing actions, in particular for its distinction between two different types of assessment at which laws of nature may determine people's behaviour. We are now able to recognise that the first type is about imputing (ascribing) personhood. Assessing an event as 'free' according to that first type of assessment (i.e., describing it as a free action in a transcendental sense) therefore means ascribing to whatever was the source of that event, the status of 'person', which consists in being the 'primary cause', the 'free cause', 'cause that is uncaused' of that action.⁴⁰

The late medieval discussions on rational determinism remained important to theorists from the early modern period and probably continued to influence discussions on necessity well into the eighteenth century. But along the way, rational determinism became more and more modelled on the vocabulary of causation due to the promising and exciting developments in theorising about the natural world.⁴¹ As a result, debates shifted from rational determinism towards psychological necessity, exploring the relations between reason, will and emotions ('passions', 'sentiments'). During the early modern period also a new notion of necessity, unknown to medieval theorists, *mechanical* necessity, was emerging, but only gradually because theorists and scientists were still struggling with the proper meaning of what it meant that *res extensa* was mechanical: was there any vacuum between atoms, how was it possible that atoms moved

37 See, e.g. Darwall (2012). This does not mean, however, that the contractarian notion of personhood was necessarily coherent. It was not. I will return to this point in the next chapter.

38 See especially Kobusch (1993 and 2006), who has tried to uncover this (what he calls) 'forgotten tradition' (2006). Also, see Hunter (2001) and Haakonssen (2010). They disagree, however, with Kobusch on what exactly is the status of the 'person': Far from attempting to provide a moral-ontological anchor for moral freedom and natural rights [as Kobusch tries to interpret Pufendorf], Pufendorf's separation of moral from natural being is intended to divorce duties and rights from all ontological foundations and salvific aspirations, grounding them instead in imposed offices which originate in a civil rather than a metaphysical order' (Hunter 2001:165-166). Also see Haakonssen (2010:1, note 3). Their point, in my view, is precisely the novel meaning of personhood as an *ascribed* status rather than a *natural* status.

39 Understanding early modern theorists as being committed to the medieval transcendental notion of personhood would explain why for example, theological determinism, as it was expounded in Calvinism, was not scandalous, although it certainly posed political problems. If anything had been scandalous in the medieval and early modern period, it would have been to deny that people, as persons, are moral entities rather than to deny that they are free.

40 Hruschka (1986).

41 One of the influential works in this area was, of course, Isaac Newton's *Philosophiæ Naturalis Principia Mathematica* (1687). It should be reminded, though, that Early modern philosophers, scientists, theorists of law and others did not have a modern notion of nature. See, e.g. Saine (1997, esp. chapter 1); Daston (1998); Daston and Stolleis (2008a); Roth (2008). This also holds for Newton, who, like other contemporaries who helped to develop the idea of a 'mechanical' world view, was still convinced that, within a mechanised world, miracles could still take place: see Harrison (1995).

and what is matter anyhow?⁴² In the eighteenth century, it began to move towards a more modern notion of mechanical determinism. Each of the two notions of necessity had its own domain of application, respectively (in line with Descartes' distinction) *res cogitans* and *res extensa* or, in other words, the domain of mind and the domain of nature.⁴³ Psychological necessity was not usually thought of as reducible to the domain of nature.⁴⁴ More importantly, neither was it understood as entailing the denial of moral responsibility.⁴⁵ Ideas of necessity, in whatever variety, were generally still underpinned by a firm belief that the moral order was secured by God. The explanation for the early modern view that necessity does not yet deny moral responsibility, apart from the theological explanation, might be theorists who were engaged in debates on necessity were aware there was a difference between agency and actions.⁴⁶ In this sense, the notion of personhood remained relevant, although very much in the background, to keep the domain of mind and the domain of nature apart. It made it possible for theorists to be necessitarians (namely, about the domain of mind) without being either compatibilists or incompatibilists in today's senses of these words.

The nineteenth century was a period of various transformations of the debate on necessity. The notion of 'necessity' finally transformed into the modern notion of 'determinism'.⁴⁷ In this process, the notion of personhood was gradually lost sight of.⁴⁸ Before, it had functioned to keep apart two distinct discussions on necessity and with it two levels of analysing freedom. Its disappearance now allowed the two discussions to converge into one discussion. The collapse of the notion of psychological necessity into the notion of

42 The early notions of mechanical necessity were still partly theological because God was introduced to explain the movements of atoms. See Dijksterhuis (1950). Also, see, e.g. Yolton (1983); Sarasohn (1985).

43 One of the early occurrences in German of the word *Determinismus* (in 1789) refers to psychological necessity, suggesting that even towards the end of the eighteenth century, this was still distinguished as a separate kind of necessity (Hacking 1983:458).

44 Harris (2014) mentions Clarke as introducing 'moral necessity' to set the mental (moral) order apart. Also, see Wokler (1995:40), who mentions that, for example, Montesquieu also acknowledges a distinction between two kinds of causes: physical and moral. The distinction between 'moral necessity' and 'physical necessity' may well have been one of the first steps in a long philosophical debate to finally make the distinction between *justification* and (*causal*) *explanation*. Rorty has claimed that seventeenth-century writers generally did not yet make this distinction because they 'did *not* think of knowledge as justified true belief' (1979/ 2009:141).

45 Harris (2014).

46 Harris (2014).

47 Hacking (1983:460): 'Somewhere, let us say between 1854 and 1872, the concept of determinism acquired its modern sense in all the European languages'. For a detailed analysis of the transition of 'necessity' to 'determinism': Cassirer ([1936] 2004). In what later would become the U.S.A., the debate on necessity was dominated, well into the nineteenth century, by Jonathan Edwards's *Freedom of the Will* ([1754] 1957), which also did not imply the impossibility of morality (Tweney 1997). Here too, during the nineteenth century, this theological notion of necessity became replaced with a naturalistic notion of necessity (Guelzo (1989).

48 See, e.g. Kobusch (1993), although he explores the decline of this notion mainly within the tradition of German philosophy. Also, see MacIntyre, who traces the turning-point in the 1780s when Kant and Reid, in their writings, began to develop a 'schism between moral philosophy on the one hand and the philosophy of mind and action on the other' (1982:311).

mechanical necessity proceeded in two steps: first, the introduction of the mind as (partly) unconscious; second, the reduction of the unconscious to the world of nature. The first step took over more than a century, in which the notion of the unconscious was introduced several times.⁴⁹ (I come back to this in more detail in the next chapter.) But an argument for psychological necessity (in today's terminology: determinism) that was reducible to physical necessity required a tighter link to empirical findings of the brain. This came only (and became only acceptable) in the twentieth century.⁵⁰ And so what once started as a discussion on rational determinism was over many centuries dissolved into a discussion on mechanical determinism. The function of hinge between the two levels of freedom, in this debate at least, shifted definitively from the notion of personhood to the notion of will.⁵¹ Whereas their status of 'moral entity' had previously functioned as what demarcates human beings from non-human beings, in the modern debate on determinism, this status was understood in terms of what I call transcendental freedom. These various shifts allowed the introduction of a claim that had not been made before and, I believe, had not been possible before because towards the end of the nineteenth century, for the first time, it was claimed that the impossibility of (transcendental!)

49 Alexander Crombie (1793) has perhaps been the first to use, in an argument on necessity, the adjective 'unconscious' (in relation to people's motives). Although the term was barely used by Crombie, Thomas Reid, in a critical comment on Crombie's book (in a letter dating from the same year, 1793), nevertheless picked out this adjective to summarise Crombie's argument as centring on 'motives we are unconscious of' (1872:87). The term 'unconscious' entered the debate on determinism only systematically after Freud's notion of the unconscious had become dominant over other notions (e.g. Hospers 1958).

50 In this respect, the ties of Freudian theory to empirical research are too weak to establish this reduction. The 'natural unconscious' is more promising: see chapter 4, section 1. Already in the early nineteenth century, phrenology was suspected to involve the notion of 'material necessity', which was rejected by Sir William Hamilton as "implicit atheism" with the consequent 'negation of the moral world" (quoted in Cantor 1975:204).

51 Ironically, in roughly the same period, the notion of the will loses ground as a founding concept in legal theory and political theory, while it continues to exist in determinism and decision theory. For legal theory in the U.S.A.: Pound (1954). For legal theory in Germany and particularly France: Ranouil (1980). For legal theory in general: Gordley (1991). For political theory: Scheffler (1992), who understands the rejection of the (personal, individual, volitional) will as the basis for political theory in terms of a transition of a preinstitutional to an institutional (i.e. non-volitional) notion of desert as the basis for theories of justice. Scheffler views Rawls as a proponent of the institutional notion of desert. Some political theorists are still committed to a preinstitutional notion of desert: see Sher (1987). In general, the will as the foundation for the social contract has lost its importance, I think, as the initial assumption that *actual consent* is required has transformed into the assumption that *hypothetical consent* is required. But I do think the shift in political philosophy during the early modern period differs in an important respect from the shift in what is now known as the philosophy of mind. In early modern social contract theory, the notion of personhood still functioned as a metaphysical notion, at least in the background, to secure/ sustain the meaning of will (e.g. Hobbes). From the nineteenth century onwards, the notion of personhood was increasingly deconstructed within the philosophy of mind, under the influence of new insights in the sciences (natural sciences, but also psychology, and later cognitive science).

freedom implies the impossibility of responsibility (and more generally: of morality).⁵² Only now the connection between freedom and determinism was forged that would later dominate debates on free will versus a determined world.

5.3. Transcendentally free but non-responsible actions

The previous section started with the claim that determinists use the wrong notion of freedom. I then suggested that we understand the notion of freedom as consisting of two levels. The brief historical overview that followed indicates that ‘personhood’ and ‘will’ provide two different options to function as the hinge between both levels. Early modern theorists, as did medieval theorists, chose the first option. For them, the boundary between human and non-human beings is secured by the status of being a person. Today’s determinists choose the second. This interpretation seems to be confirmed already by a mere glance at their numerous articles and books whose titles carry ‘free will’. Their approaches to ‘free will’ have roughly the same structure. The ‘free will’ is the *explanandum*. The will is what is supposed to be made free or, to phrase it with a familiar political metaphor, to be ‘liberated’. What needs to be explained then is *how* the will is liberated by what might be called some sort of ‘liberating cause’.⁵³ The will also functions as a ‘hinge’ because it is supposed to secure both responsibility and transcendental freedom or, to put it in terms of freedom as a two-level notion, to connect the moral-political and the metaphysical level of analysing freedom: if responsibility presupposes (transcendental) freedom, the impossibility of freedom implies the impossibility of responsibility. But choosing the will as ‘hinge’ is choosing the wrong notion of freedom. To show why this is problematic, I will turn again to a discussion of Frankfurt’s account of free will in ‘Freedom of the Will and the Concept of a Person’.

Frankfurt’s account is illustrative of an important and dominant category of compatibilist views that are sometimes called ‘Real Self’ views.⁵⁴ The problems with choosing the will rather than personhood as ‘hinge’ are apparent most in these views because these can be understood as trying to connect the notions of ‘will’ and ‘person’. Of course, they use the will as a hinge, but they attempt to trace the origins of its freedom – the ‘liberating cause’ – in some modernised notion of ‘person’, more specifically as: ‘personal identity’, ‘personal self’, ‘deep self’, ‘true self’, to name just a few terms that are used.⁵⁵ Their choice

52 See McKeon (1957), who shows how the debates on responsibility and free will initially were distinct debates but became linked by John Stuart Mill. Later, Bradley [1876] (2006) is perhaps the first one in the history of modern philosophy, who explicitly claims – within a context that is clearly not moral-political, but transcendental (in my sense) – that freedom precedes morality. For a recent restatement of the impossibility of moral responsibility: see Strawson, G. (1993).

53 Cf. what Kershnar (2015) calls the ‘responsibility-foundation’.

54 The phrase ‘real self’ was introduced by Wolf (1990, chapter 2).

55 I think it is no coincidence that they are closely related to so-called autonomy views. They both understand freedom in terms of personal ownership, either as owning one’s actions or owning one’s identity (as the source of freedom of action).

of the term ‘self’ rather than ‘person’ (which only shows up as an adjective to ‘self’) has a reason; I think because they primarily deal with the issue of transcendental freedom and usually ignore the demarcation issue that is associated with the notion of personhood. For this reason, Frankfurt’s account is especially interesting to our discussion because it presents a rare, modern example of an account that tries to combine the notions of personhood and the will in the sense of combining the demarcation issue and the freedom issue. Frankfurt discusses not only how the will’s freedom originates in the ‘Real Self’ but also the differences between persons and non-persons. The problems with connecting the notions of ‘will’ and ‘person’ in the way ‘Real Self’ views do comes more to the surface in Frankfurt’s account than in other ‘Real Self’ views.⁵⁶

Frankfurt’s account is usually referred to as a proposal to explain ‘free will’. What is usually not commented on is that his article starts off as an analysis of *personhood*, only to move on in later sections to discussing *freedom of the will*.⁵⁷ Frankfurt’s article originates in a criticism of Strawson (1958 and 1959, esp. chapter 3), whose notion of personhood he thinks is too inclusive: it covers not only human beings but also ‘animals of various lesser species’ (Frankfurt 1971:5). Strawson’s analysis of personhood is part of a distinctively modern contribution to the conceptual history of ‘personhood’. Here ‘person’ has survived as a technical term but has lost its original, medieval meaning of the person as a moral entity, although it still seems to preserve the idea of the person as a rational entity. Instead, it is used to analyse the problems of identity and the unity of body and mind.⁵⁸ In this modern tradition, personhood, in this sense, is not usually linked up with the (transcendental) problem of free will. Frankfurt acknowledges that Strawson is primarily interested in ‘the problem of understanding the relation between mind and body’. For Frankfurt, this is not enough. The problem of understanding personhood is not just about understanding the relation between mind and body, but about ‘understanding what it is to be a creature that not only has a mind and a body but is also a person’ (1971:5, note 1). And so, what Frankfurt has in mind in criticising Strawson, and the tradition he represents, seems to be no less than arguing for a paradigm shift in the debate on personhood:

‘the criteria for being a person [...] are designed to capture those attributes which are the subject of our most humane concern with ourselves and the source of what we regard as most important and most problematical in our lives’ (1971:6).

56 Wolf (1990) herself mentions Frankfurt’s account as illustrative of ‘Real Self’ views.

57 Frankfurt’s account (1971) is usually discussed *only* as providing a compatibilist notion of freedom. Only a few have commented on the fact that it is also an account of personhood: e.g. Morton (1990), Ausborn-Brinker (1999) and Hermann (2001). The title Frankfurt himself gave to his article actually reverses the order in which he treats his two themes.

58 See Ayer (1963); Dennett ([1976] 1978); Parfit (1984); Carruthers (1986). For a critical discussion of several key theorists in this tradition: McCall (1990) and Hermann (2001).

These phrases seem to suggest that Frankfurt is perhaps interested in reintroducing a moral dimension to personhood or else to explain the boundary between persons and non-persons as one between people and animals because, as Frankfurt claims in his criticism of Strawson, it ‘does violence to our language to endorse the application of the term ‘person’ to those numerous creatures which do have both psychological and

material properties but which are manifestly not persons in any normal sense of the word’ (1971:5). As we have seen in chapter three, what distinguishes a person from a non-person, i.e. a ‘wanton’, is having second-order desires or, what amounts to the same, a person’s individual identity: her ‘real self’. A person can be free because her second-order desires (her identity, her ‘liberating cause’) allow her to select the appropriate desire (i.e. the desire she can identify with) and turn it into an effective desire: free will.

Once the will’s freedom is traced to the Real Self, an obvious question follows: should not the Real Self also have transcendental freedom? The problem with second-order desires (or any other version of the Real Self) is that they may be understood as part of the natural chain of causes. Second-order desires may therefore constitute a ‘mechanism’ themselves. It is not clear then how a ‘liberating mechanism’, itself subject to mechanical laws, might be able to ‘liberate’ the will in a sense that somehow withdraws the will from mechanical laws. Roughly the question may be answered in two different ways: the ‘liberating mechanism’ is liberated itself by yet another ‘higher-order mental cause’, or it liberates itself; in other words, it is a *causa sui*. In both cases, the ‘liberating process’ is supposed to be realised in a certain process of reflection (whatever extra constraints are imposed on this process). Both responses have been criticised, either by showing the incorrectness of the underlying principle of *causa sui* or by claiming that the supposedly ‘liberating’ process of reflection presupposes an infinite regress.⁵⁹ Both criticisms are correct, I think. But that may not yet be the real problem because, whatever the responses to such worries, they do not lead to a radical reframing of the debate, for example, by shifting from the will towards the person as central notion and simultaneously treating the freedom issue in close connection to the demarcation issue. They do lead to new, ever more complex and nuanced explanations of how the will can be free, but only by ignoring a more fundamental notion: ‘person’. Let us return to Frankfurt’s account as an example of how the issue of personhood is really ignored.

In the previous chapter, I presented one possible interpretation of Frankfurt’s account: one that takes Frankfurt to claim that transcendental freedom is only freedom of the mind, an ‘inner citadel’. But his account allows still other interpretations. The standard interpretation is one that takes Frankfurt to claim freedom of action as having its origin in an explanation to which the distinction between persons and wantons is essential. Here I will argue for an interpretation that denies the importance of what persons are in distinction to wantons.

59 See, e.g. Galen Strawson (1993), Kornblith (2012, chapter 3.1).

His notion of second-order volitions functions to mark the difference between persons and non-persons. But this demarcation does not have a moral significance: 'I do not mean to suggest that a person's second-order volitions necessarily manifest a moral stance on his part toward his first-order desires' (1971:13, note 6). Neither does it

support the claim that people are, as a species, essentially different from non-persons.⁶⁰ As I have stated before, medieval and early modern theorists believed that persons are different from non-persons because they are moral entities. For today's theorists, this is no longer an acceptable idea. Instead, some of them believe that persons are different from non-persons for having free will. Consequently, if some entities can be understood to have free will, they are persons. Also, if no entity can be understood to have one, as hard determinists believe, persons do not exist. People do exist, of course, but only as particulars of a biological species. If freedom constitutes the boundary between persons and non-persons (i.e., wantons, animals, children perhaps), Frankfurt's analysis is a muddled one. His notion of free will draws this boundary, not his notion of personhood. His aim of analysing the conditions of free will fits in with the determinist debate. But what about his aim to analyse the conditions of personhood? Being a person, i.e. having second-order desires, is a necessary condition for having free will, but not a sufficient one. Being a person is no guarantee of having free will. Frankfurt's account has been enormously influential, but only in his proposal for analysing 'free will', not in disconnecting 'having a free will' and 'being a person' (which in determinists' readings of Frankfurt are usually, uncritically, thought to be linked). Frankfurt's notion of personhood leaves room for a limbo where entities reside that neither are wantons nor have a free will. But what exactly is their status? If they do not have free will, there is no reason to suppose that what distinguishes them from wantons is more special than what distinguishes wantons from beings that have no consciousness. Unfree persons (such as unwilling addicts), wantons and non-conscious entities all have in common that they are subject to mechanical laws. Frankfurt's aim to capture, in his notion of personhood, 'those attributes which are the subject of our most humane concern with ourselves' therefore fails. Frankfurt's account of transcendently free beings is therefore too exclusive. And it ignores, as do other determinists, that the 'person' itself might be a (better) candidate as the bearer of transcendental freedom.

The reintroduction of the notion of 'person', not as what explains the will's freedom, but as what itself has transcendental freedom, opens an interesting option that is usually overlooked in the debate on determinism. Consider once again Pereboom's case of Professor Plum. It is part of his strategy to make a case for hard determinism (chapter 4, section 2).

⁶⁰ In the last paragraphs of his article, Frankfurt even concedes the possibility that 'it is determined, ineluctably and by forces beyond their control, that certain people have free wills and others do not' (1971:20). At this point, Frankfurt's view surprisingly moves close to Hobbes's view on the reconcilability of the two types of freedom.

The reasoning behind it might be something like this. Responsibility is defined by (some) compatibilists in terms of a 'will' condition *and* an 'effectivity' condition. Professor Plum is responsible for murdering Ms White if he wants to kill her (the 'will' condition) and if his will to kill her really is the actual cause of her death (the effectivity condition).⁶¹ We

might add a 'knowledge' condition: Professor Plum *knows* that his will to kill her is the actual cause of her death. This provides the key to understanding the hard determinist strategy. Now, if responsibility presupposes freedom, as all determinists seem to agree on, for Pereboom, the first step in his strategy is to make plausible how Professor Plum might *believe* he is the actual cause of the murder while the actual cause is really something else: some form of manipulation. Crucial here is that the cognitive status of what Professor Plum thinks about his murdering act changes from 'knows X' to 'believes X'. The second step then is to generalise this situation to all situations in which Professor Plum (or, for that matter: all people) believes he is the actual cause of some action. From here, it is only a small step to accepting that the world is deterministic (in the way that hard determinists understand it). If you accept one situation of manipulation, why not accept that all situations might be cases of manipulation?

This reasoning would mean that two options are available for assessing Professor Plum's responsibility for murdering Ms White:

1. Professor Plum *is responsible for* murdering Ms White because, next to fulfilling certain responsibility conditions, he has the transcendental freedom that is required for responsibility (the compatibilist view).
2. Professor Plum *is not responsible for* murdering her because he has no transcendental freedom, i.e. because he lives in a mechanical world (the hard determinist view).

But still, a third option is possible:

3. Professor Plum *is not responsible for* action X, and yet he has transcendental freedom.

This option comes into view once we accept that personhood rather than the will might be the bearer of transcendental freedom. It would allow a more complex assessment, as in the case of Pufendorf's view on imputation, for example, by claiming that someone is a person (and therefore has transcendental freedom) and yet is not responsible for what she did. Such actions belong within the category of 'excusable actions'. Pereboom's

61 This corresponds to what Frankfurt distinguishes as respectively 'freedom of the will' and 'freedom of action' (also see chapter three, section 2o). More on the relation between Frankfurt's two notions of freedom: this chapter, section 7.

binary view, however, does not allow him to distinguish between a situation in which Professor Plum is always free and responsible for his actions (because manipulation does not exist), a situation in which he is basically unfree and never responsible for his actions (because he is ‘manipulated’ all the time) and a situation in which he is basically free as a person and only sometimes not responsible for his actions (perhaps because he was manipulated, perhaps for some other reason). The reason for this is that Pereboom, and most of the determinists, tend to conflate the two levels of analysing freedom, i.e. the distinction between unfree and free actions and the distinction between responsible and excusable actions.⁶² Responsible actions are aligned with free actions, while excusable actions are generalised to coincide with unfree (determined) actions.⁶³ If Professor Plum was judged according to a juridical model of responsibility, the conclusion might be that he is not responsible for the death of Ms White. After all, if it were assessed that some evil doctors had manipulated Professor Plum, this would exonerate Professor Plum. (And of course, hard determinists cannot argue that Professor Plum is nevertheless guilty because he would have killed Ms White if the manipulation had not occurred because the manipulation was determined, was it not?) And again, according to this model, it is not evident that having no responsibility implies having no transcendental freedom. If this option is a plausible one – not being responsible and yet having transcendental freedom –, it would show that the story about Professor Plum is inadequate for making the hard determinist view plausible.

5.4. Brandom’s radical position

If I am right that determinists make a category mistake and have singled out the wrong notion of freedom – the ‘will’ instead of ‘personhood’ –, this does not yet provide an argument against determinism. But at least it invites an exploration of freedom that takes ‘person’ as a starting point.⁶⁴ This notion should help to explain how someone,

- 62 Hruschka (1986, esp. Part XIV) and Smiley (1992) have argued for a similar point. The distinction between two different levels of responsibility – transcendental versus moral-political, first-level imputation versus second-level imputation (Hruschka 1986), causal responsibility versus blameworthiness (Smiley (1992) – allows for a more nuanced analysis of blame, merit and excusability. For Smiley’s critical analysis of modern theories of moral responsibility that ignore this distinction: Smiley (1992, chapter 5). Smiley does not start from a two-level notion of freedom as explicitly I introduced in this section, but she does claim that two levels are conflated that more or less agree with the two levels I distinguish.
- 63 The conflation of levels of analysing freedom can be recognised in the way the language of ‘excuses’ is sometimes used within determinist debates. For example, it is asked whether the fact that people are not free but determined may function as an excuse. But using the language of ‘excuses’ is unnecessarily confusing matters. Excuses, in the tradition of Pufendorf and therefore in the tradition of criminal law that followed it, only make sense when we recognise the one to whom the action/event is recognised as a person. This is exactly the problem in determinist debates: it denies the existence of persons (in the sense of ‘moral entity’).
- 64 In the naturalistic debate on free will, such proposals are said to be committed to ‘agent causation’ and are sometimes called ‘agent-causal libertarianism’. Only quite recently is it being discussed. Proponents are, for example, Hasker (1999), O’Connor (2000), Clarke (2003). O’Connor mentions as early predecessors Thomas Reid. He does not mention Kant. More recent research has shown, however, how Kant also fits in this line of thought: Watkins (2005). Also: Blöser (2014).

as a person, performs acts that are sometimes responsible and sometimes excusable and still be transcendently free in both cases. The strategy that will be pursued here is to understand personhood in terms of rationality: the person as a rational being.⁶⁵ In this and the next sections, I will discuss Robert Brandom's account of freedom. It chooses rationality as a way of drawing the boundary between human and non-human beings.⁶⁶ Basically, people are 'claim-makers', which we may use as a shorthand phrase for how Brandom understands people as rational, *sapient* beings.⁶⁷ At the heart of this idea of people as claim-makers is the notion of responsibility: to be responsible for what we claim is what makes us free. Not the other way around. Brandom's position in connecting rationality to freedom and responsibility differs from the standard positions within the determinist debate in several respects.⁶⁸ To show how radical his position really is within this debate, I will indicate what these differences are, but only at an abstract level. The details of the argument will be provided over the course of the next sections.

To those who believe that freedom and responsibility are compatible with determinism, basically, two grand strategies are available: the Real Self Views, to which Frankfurt's account belongs, and the Reasons-Responsiveness Views.⁶⁹ The most influential version from this latter category is probably Fischer and Ravizza's account (1998).⁷⁰ The difference between the first and second strategies is *not* that the second relies on rationality, whereas the first does not.⁷¹

65 Claiming a link between freedom and rationality is a familiar strategy of arguing that human beings are, in principle, different from non-human beings. In a sense, it is a return to the Aristotelian strategy of characterising the human being as animal rationale.

66 Of course, the boundary between human beings and animals may be drawn in yet different ways. See, e.g. Shalamov (1980), who writes, in one of his stories, based on his own experiences, on the Soviet camps in Siberia: 'A horse weakens and falls ill much quicker than a human being [...] A horse can't endure even a month of the local winter in a cold stall if it's worked hard hours in subzero weather. [...] But man lives on. Perhaps he lives by virtue of his hopes? But he doesn't have any hope. He is saved by a drive for self-preservation, a tenacious clinging to life, a physical tenacity to which his entire consciousness is subordinated. He lives on the same things as a bird or dog, but he clings more strongly to life than they do. His is a greater endurance than that of any animal.'

67 For explaining his view on human rationality, Brandom does not use the phrase 'claim-maker', but he might agree with it, I think. I was drawn to this phrase by a remark made by Rawls (1980:543) when he said that 'citizens think of themselves of as self-originating sources of valid claims' (also see my discussion in section 1.5.). While Brandom's and Rawls's views on people as claim-makers indicate interesting similarities, especially in linking freedom and rationality, what makes them different, I would say, is that Brandom's notion functions at the metaphysical level of freedom while Rawls's notion functions only at the moral-political level.

68 Brandom's non-naturalist position is not yet clearly recognised in current debates. Baker, for example, mentions Brandom as one of the 'nonscientific naturalists' (2013:26). I do not think this characterisation does justice to Brandom's position, as will become clear in this chapter.

69 This idea of rationality is associated by some participants in the debate with classic philosophers such as Aristotle (Irwin 1980; McKenna 2013; Hurley 2003:32, note 13) and Kant (McKenna 2017; Hurley 2003:32), Spinoza and Hegel (Hurley 2003:32).

70 One of the most prominent, and perhaps also most detailed, view in the determinist debate itself is the one presented by Fischer and his co-author Ravizza. What is characteristic about their version is that it is also committed to the ideal of individuality as a feature of freedom.

71 For example, Ausborn-Brinker (1999:126-128) claims that Frankfurt's analysis of the structure of their will that is specific to persons should be understood really as an analysis of their rationality. For a similar view: Herrmann (2001:173-177).

In fact, they both rely on the notion of deliberative rationality (they do not use this notion in exactly the same way: I will come back to this in the next paragraph). And both strategies assume, in the standard views at least, that the issue of freedom and responsibility turns on the link between freedom and instrumental rationality (see section 1). Here we meet the first difference between these two strategies and Brandom's own strategy because Brandom, as I will explain later in this chapter, links freedom to theoretical rationality rather than instrumental rationality.

The second difference relates to *how* theoretical rationality, once it is conceptualised as the source of freedom, is then linked to instrumental rationality. In the previous chapter, we have seen how some philosophers nevertheless explore the issue of having free beliefs or, as we will frame it here, of theoretical rationality. Brandom would also reject their approach, I think. Although they do not ignore theoretical rationality, they explore its relevance for freedom and responsibility only to the extent that it contributes to instrumental rationality.⁷² The theoretical problem this leads to is illustrated by some of the deliberations on this subject by Isaiah Berlin, who, as an outsider to the debate, addresses determinism in his 'From hope and fear set free' (1963/1964). For Berlin, freedom appears to be, surprisingly, a matter of 'positive freedom' because in defining transcendental freedom, he singles out the very feature that he rejects for the moral-political notion: 'self-direction' (which seems only another word for self-mastery).⁷³ This need not be problematic. After all, he addresses here, in my view, another type of freedom. The clue to understanding his definition is the question that starts his discussion on determinism: 'Does knowledge always liberate?' Although knowledge and rationality are not necessarily identical, Berlin does intend to ask about the sense in which rationality is liberating. He follows the Enlightenment idea that to be rational is also to be rational about how one acquires knowledge and that this rational process of knowledge acquisition may set us free. Interestingly he uses the state of being drugged or hypnotised as a metaphor for the loss of transcendental freedom. Berlin is opposed to determinism.⁷⁴ He nevertheless raises doubts about whether rationality will really set us 'beyond hope and fear' because, assuming that reality is 'a rational whole (whatever this may mean)', in the end 'the discovery of its structure will not increase my freedom of choice' (Berlin 1999:196 [1963–1964]). He wonders:

'is all liberty just that? The advance of knowledge stops men from wasting their resources upon delusive projects. It has stopped us from burning witches or flogging lunatics or predicting the future by listening to oracles or looking at the entrails of birds. [...] It may make our conduct more rational, perhaps more tolerant, charitable, civilised, it may improve it in many ways, but will it increase the area of free choice? For individuals or groups?' (1963–1964:28).

72 I think this holds, at least for both Korsgaard and Moran, as representatives of the Real Self View. It is less obvious in the case of, for example, Arpaly, as a representative of the Reasons-Responsive View (also see my remarks in this chapter, note 78).

73 Berlin (1963-1964:173).

74 Berlin seems to be one of few to have observed that even determinists remain committed to categories of morality: see Galipeau (1994:71-80).

Berlin connects his transcendental to his moral-political notion of freedom: he wants to capture how rationality, in the sense of the acquisition of ever more knowledge, contributes to people's free choice. And exactly there, as the quote above indicates, he loses grip on how his notion of transcendental freedom may pose a good answer to determinism. The problem is that Berlin, while he turns the acquisition of knowledge into a function of moral-political freedom ('free choice'), treats instrumental rationality as somehow prior to theoretical rationality. As we will see in this chapter, Brandom instead treats theoretical rationality as prior to instrumental rationality. Even more so, Brandom treats theoretical rationality itself as practical rationality, the genus of which instrumental rationality is only the species.⁷⁵ This is Brandom's pragmatism about human cognition, and it represents a position that is, I think, largely absent in the determinism debate.

Even if we disregard the previous difference, still a third difference may be identified on how theoretical rationality is conceptualised. This requires some explanation about how Real Self Views and Reasons-responsiveness views differ from each other in their notion of theoretical rationality. Those few compatibilists who do discuss theoretical rationality in the context of the determinism debate assume that it turns on the question of whether people have control over their beliefs or not. The two positions that are available here, voluntarism or involuntarism, to some extent run parallel with one specific difference between Real Self and Reasons-responsiveness views.⁷⁶ The first start from a concern with understanding what it means that reasons are one's own. The second usually start with rationality and then, in some cases, go on to explore what it means that reasons are someone's own. This has relevance at a deeper level. Adherents of Real Self Views have an interest in securing, conceptually, the 'own-ness' of reasons. The standard strategy in (early modern) philosophy has therefore been to trace this 'own-ness' to some process of reflectivity (i.e. *introspection*; also see chapters two and three). In the previous chapter, we have seen what challenge this poses to Real Self Views. Adherents of reasons-responsiveness views have tried to tackle this problem by shifting from a reflective to a non-reflective notion of deliberative rationality.⁷⁷ So whereas proponents of Real Self Views rely on a notion of rationality that is closely linked to reflectiveness (transitional

75 Berlin (1963-1964) struggled with the question of whether knowledge (rationality) is a liberating force against a deterministic universe. In the end, he does not manage to solve this issue because he is not able to see the option Brandom suggests of how theoretical rationality (what Berlin calls 'knowledge of facts') may also be understood as a form of practical rationality (what Berlin calls 'knowledge of "what to do"').

76 To my knowledge, there are only a few efforts of comparing both types of views (see Sripada 2015), let alone in terms of how they use the notion of rationality. Admittedly the boundaries that I draw up between them in these paragraphs are open to discussion. Concise as my argument is, it is meant in the first place to give a rough idea of how radically Brandom's account of rationality differs from standard approaches in the determinism debate.

77 This is a development that is usually ignored within the 'Rationality Wars'. For a review of the psychological literature on the purported failure of human rationality, however, without any reference to this philosophical discussion: see Rysiew (2008).

consciousness), proponents of Reasons-Responsiveness views deny that reflectiveness is essential to reasons-responsiveness.⁷⁸ This brings us back to a discussion that was briefly mentioned in the previous chapter (section 3): one between deontological voluntarists and deontological *inv*oluntarists. It would mean, in my view, that adherents of reasons-responsiveness views belong to the second group.

Brandom would reject the basically reflective notion of rationality. But would he agree with the involuntarist position about beliefs? For medieval philosophers, as we have seen in section two, the choice between being a voluntarist or an involuntarist (i.e. an intellectualist) was not a matter of being a soft or hard determinist. Instead, their discussion was probably, more than anything else, an effort to understand more exactly what rationality was about and how it worked:

‘Debates [between intellectualists and voluntarists] about whether the will can act against intellect’s judgment, for example, usually mask deeper differences about what intellect’s “judgment” amounts to and how it arrives at its judgment in the first place’ (Kent 1995:116).

In the modern period, the discussion about the acquisition of knowledge also became tied, as did the discussion about responsible actions, to the question of transcendental freedom. Historically two main approaches can be distinguished: the Cartesian, voluntarist and the Humean, involuntarist approach. According to the first, people articulate their ‘knowledge in voluntary judgements whereby we either assent to or dissent from some relevant mental representation. Thus, each item of conscious knowledge is gained through an act of free-will, and any thought that is admitted or accepted may also be rejected if there is reason to doubt it’ (Cohen 1992:1–2). Humeans, on the other hand, ‘treat cognition as the domain of causal laws that explain how our mental feelings originate or how one passive state produces another’ (Cohen 1992:2). Interestingly the strategic move to disconnect rationality from reflective consciousness (i.e. introspection) seems to run parallel, for some philosophers at least, with yet another strategic move, namely to disconnect normativity from freedom.⁷⁹ It allows philosophers like Arpaly (2006) to be compatibilists about freedom of action but to deny that people have control

78 For example, Fischer and Ravizza remark: ‘the process [of taking responsibility, i.e. of developing into a reasons-responsive being] need not be explicit, conscious, or reflective’ (1998:243).

79 See Owens (2000), Arpaly (2006), Sher (2009).

over their beliefs.⁸⁰ Although their shift from a reflective to a non-reflective notion of rationality has the advantage of making their views invulnerable to the Manipulation Argument, this obfuscates a problematic tendency in these views, namely to understand rationality ('reasons-responsiveness') as a natural mechanism and therefore as reducible to mechanical necessity. Exemplary for this is how Fischer and Ravizza, in their influential account, explain reasons-responsiveness in terms of a mechanism (1998:38–41). This is more than a metaphor because they relocate reasons-responsiveness from the agent towards that mechanism. It brings, however, their account close to a naturalistic explanation of rationality.⁸¹ Brandom agrees that cognition is about bindingness, but he would disagree with Arpaly's strategy to accept this bindingness as unfreedom and instead remain committed to normativity. And he would also disagree with Fischer and Ravizza's strategy to 'mechanise' rationality. Instead, Brandom chooses to understand freedom in terms of bindingness rather than the will (choice, control, decision). Rather than choosing between voluntarism or involuntarism, freedom or normativity, Brandom chooses to understand freedom as rationality, rationality as bindingness and bindingness as a non-mechanical, i.e. normative, process.

5.5. Rationality as normative bindingness

A chunk of iron, a thermometer, a parrot and a human being: they all can be interpreted – in what they *do* in the interaction with their environment – as having a *reliable differential responsive disposition* (in short: an RDRD). A chunk of iron will respond to a wet environment by rusting, to another environment by melting, to yet another environment by falling. A parrot may be trained to utter 'that's red' when seeing red objects.⁸² And of course, we may also hear a human being say 'that's red' when seeing a red object. So far, Brandom and naturalists would agree on objects, animals and people as having an RDRD. Additionally, naturalists would hold that in each case, the differential response can be said to be governed by the mechanical rules of nature. Naturalists like Carruthers (2013b) would nevertheless agree with Brandom that objects

80 Traditionally the distinction between rational and non-rational has been treated as running parallel with the distinction between conscious and non-(or: un-)conscious. Theorists such as Owens and Arpaly have the ambition of defending a notion of 'being rational' that extends beyond 'being conscious', i.e. also covers situations in which people do not seem to be conscious of their rationality. Once this strategy is linked, however, to the strategy of denying normativity, either by denying that 'ought' implies 'can' (e.g. Arpaly 2006, chapter 3) or by endorsing the view of deontological involuntarism (e.g. Arpaly 2006:91-101), they are not for removed, in my view, from hard determinists such as Pereboom who deny freedom, but remain committed to morality. Arpaly remains committed to a notion of (personal) responsibility (see her discussion of Pereboom's story of Professor Plum; 2006:85; also see 2006:37-39), but I fail to see how these various strategies fit together (*pro* responsibility, *pro* determinism, *pro* rationality, *contra* normativity, *contra* modality).

81 Fischer and Ravizza (1998) have been criticised for this strategic move on several grounds. For a brief discussion of these criticisms: see McKenna (2013:160-163). One criticism that comes close to mine is Wallace's.

82 The chunk of iron and the parrot belong to Brandom's favoured examples (e.g. 1994:88 and 2000:48).

on the one hand and sentient beings (animals and people) on the other hand respond to their environment in different ways. Sentient beings differ from objects, chunks of iron, for example, in having *intentionality*.⁸³ The behaviour they exhibit is not just a matter of stimulus-response. They can be said to have beliefs that intermediate between certain stimuli and the responses they ultimately give. This is still in agreement with the functionalist view we saw in the preceding chapter. And probably Brandom would still agree with naturalists. But he would also claim that people are both sentient and *sapient* beings. What distinguishes them from merely sentient beings is that their intentionality is of 'the fanciest sort': discursive intentionality.⁸⁴ Articulating what discursive intentionality consists in, as what distinguishes human from non-human beings, is one of Brandom's main philosophical concerns. His strategy is to explain it in terms of normative rather than causal-mechanical bindingness. Part of his strategy are two moves: the shift from epistemology to semantics and, next, the shift from semantics to pragmatics. The second will be explored in the next section. The first is discussed in this section.

In exploring the demarcation issue, Brandom's writings are largely concerned with issues of language and logic. But running through them is Brandom's opposition against attempts at naturalising human sapience (rationality). And yet, Brandom has not been engaged directly in debates on free will. Only rarely does he comment on it. He did once, however. In an interview, he was asked his opinion, against the background of advances in cognitive science, on the debate on freedom of the will. He responded like this:

'The pre-Kantian philosophical tradition thought that the notion of choice was prior to the notion of responsibility in the order of explanation. That is, it was part of the rising tide of modernity to think of people as responsible only for what they had chosen to do. And the question of free will, of what one had or even could choose to do, what one could do otherwise, accordingly came to the fore. But I take it that it is an essential part of the Kantian conceptual revolution to reverse that order of explanation. Freedom is a matter of responsibility and authority. It is a normative status for Kant, not a matter of factual status, turning on the question of whether one could have done otherwise. Freedom is, for Kant and for Hegel, the capacity to commit oneself, to be a subject of normative statuses, to exert a certain kind of authority' (Pritzlaff 2008:376).

It is not immediately clear, from the interview or from his writings, what Brandom means by *reversing* the order of explanation between choice (i.e. authority) and responsibility. At least he is hinting here at an order that was central to the early modern, social contract

83 Also, see chapter 2, notes 119 and 120.

84 Because of this, our sentience, too, differs from merely sentient beings (see Brandom 1994:276).

tradition. Its theorists conceived of freedom as being a voluntarist (in the sense of pre-normative, pre-moral) and individual achievement.⁸⁵ In my view, Brandom means to turn both features into their opposite: freedom would essentially be a normative *and* social achievement. For now, I will deal with the first feature; the second feature will be discussed in the next chapter.

The social contract tradition is nevertheless far removed from explicit concerns with determinism or the demarcation issue, as I suggested earlier (section 2). Theorists within the early modern tradition were primarily interested in analysing actions at the level of moral-political freedom. At the same time, despite his critical though appreciative assessment of the social contract tradition, Brandom himself is more concerned with transcendental than with moral-political freedom.⁸⁶ The reason, however, for picking up on the social contract tradition is that for Brandom, transcendental freedom is intimately connected with moral-political freedom, notably with Kant's notion of autonomy.⁸⁷ Earlier I suggested (chapter 2, section 2) that in his view, Kant's notion of autonomy solves an important problem within the social contract tradition – the puzzle of responsibility as heteronomy – because it consists in authority *and* responsibility: Kant's basic idea is that *taking authority* for one's moral norms by applying the categorical imperative to one's maxims means making oneself responsible for whatever is the outcome of this procedure to do.⁸⁸ Brandom believes that Kant's notion somehow provides the model for transcendental freedom.

Several neo-Kantians, especially within the Anglo-American tradition, have also claimed that Kant's categorical imperative is relevant for the discussion on transcendental freedom. They have tended to favour a strategy of claiming a direct relevance of the categorical imperative for transcendental freedom: applying the moral law to the domain of moral-political actions is what by itself secures people's transcendental freedom.⁸⁹ Brandom, however, understands the intimate connection between transcendental and moral-political freedom in a different way. For Neo-Kantians, the moral law functions as a principle for the *moral-practical* domain (i.e. for deciding what actions are right or wrong). However, Brandom does not believe that the moral law may also function to secure transcendental freedom. Instead, Brandom's use of Kant's notion of autonomy seems to suggest that he transfers the idea of 'taking authority', which belongs to the domain of moral-political rationality and consists in self-bindingness, to the level of

85 We should be reminded here also that in the pre-modern period, 'morality' was understood in a broader sense than we do today. See Hruschka (1986:670, n. 1): 'Pufendorf, for example, used the adjective "moral" generally to describe human action'.

86 And so I think that Sieckmann (2012:3) is missing the point in suggesting (and criticising) that Brandom analyses, following Kant, autonomy at the moral-political level.

87 This is just one way of introducing his favoured notion of transcendental freedom. Elsewhere he suggests a similar move was made by the American pragmatists (2004).

88 Brandom is also influenced by Hegel's critique of Kant (Brandom 2002, chapter 7).

89 On the non-metaphysical perspective that Anglo-American Neo-Kantians have adopted: see chapter 2, note 54.

theoretical rationality in the sense that taking authority for one's concepts – binding oneself to certain concepts – makes up the core of theoretical rationality. Having theoretical rationality would therefore be identical to having transcendental freedom or, as Kant calls it, spontaneity.⁹⁰ He also follows Kant in his claim that spontaneity – 'the capacity to deploy concepts' (2009:59) – defines the boundary between rational and non-rational beings. Whereas non-rational beings are compelled immediately by the laws (*Gesetze*) dictated by nature, human beings are not, at least not to the extent that, as Brandom summarises Kant's core idea, 'we act according to our *grasp* or *understanding* of rules' (MIE:31).⁹¹ The term 'rules' (as a translation of 'Gesetze') belongs to two different language games, one that belongs to the natural sciences, another that belongs to jurisprudence. Brandom claims that the second is decisive for Kant's characterisation of spontaneity: 'Drawing on a jurisprudential tradition that includes Grotius, Pufendorf, and Crusius, Kant talks about norms in the form of *rules*' (2006:56). He also talks about them as 'concepts' (*Vorstellungen*). This is another way of saying that discursive intentionality basically involves conceptual activity. Whatever beliefs merely sentient beings may be claimed to have, the beliefs human beings have, are different: they are binding in a normative rather than a causal-mechanical sense. They can function as reasons, both theoretically, as judgements or claims that give reason to adopt yet other beliefs, and practically, as what motivates to perform (or not) a certain action.⁹²

Brandom's Kantian strategy would seem to promise no progress, however, beyond the discussions in the previous chapter.⁹³ Shifting from the domain of practical rationality,

90 For Brandom's discussion of Kant's spontaneity: see Brandom (1994:50-52; 2006:60-61; 2009:58-63).

91 Brandom is hinting at this passage in ([1785/1786] 1983:41, BA 36): 'Ein jedes Ding der Natur wirkt nach Gesetzen. Nur ein vernünftiges Wesen hat das Vermögen, *nach der Vorstellung* der Gesetze, d.i. nach Prinzipien, zu handeln, oder einen Willen. Da zur Ableitung der Handlungen von Gesetzen *Vernunft* erfordert wird, so ist der Wille nichts anders, als praktische Vernunft.' As Brandom explains: 'Kant initiates a shift in attention from *ontological* questions (understanding the difference between two sorts of fact: physical facts and mental facts) to *deontological* ones (understanding the difference between facts and norms, or between description and prescription)' (2002:212).

92 In a different context, Brandom remarks that what is at issue here is not complexity as such, but 'just what sort of complexity' (2009:183).

93 Pippin writes approvingly about Brandom's attempt to appropriate the 'deeply Kantian position on normativity'. At the same he is sceptical: '[binding oneself] is a *highly* metaphorical notion in all three thinkers [i.e. Kant, Fichte and Hegel]; there is no original moment of self-obligation, any more than there is a Fichtean I which initiates experience *de novo* by positing a not-I' (2007:163). A similar concern is expressed by McDowell (also quoted by Pippin), who is critical about the Kantian idea of self-legislation: 'It makes no sense to picture an act that brings norms into existence out of a normative void' (McDowell 2002:276). I think these worries can be solved within Brandom's normative pragmatics by making certain theoretical moves. Brandom himself suggests two of these moves: first by understanding bindingness as practical (see the next section), second by understanding practical bindingness as a social affair, as something we learn in the company of others. In my view, this kind of bindingness (if we follow Brandom's moves) only leads to collective commitments (see my proposal in chapter 6, section 5), not to bindingness in the sense of our own bindingness: as binding myself. What does indicate an original binding-myself, as I will argue later on (in chapter 7, section 3), is, first of all, *unbinding* oneself, i.e. disagreeing, deauthorising one's former collective commitments. This opens the possibility, next, for binding oneself.

which neo-Kantians focus on, to the domain of theoretical rationality, which is Brandom's concern, seems a reiteration of an insight we already achieved in the previous chapter, which is that transcendental *practical* freedom presupposes transcendental *theoretical* freedom. Second, one important criticism of the idea of autonomy as self-bindingness as a solution for the moral-political domain has been that it raises what has been called the *Kantian paradox*:

'How can we *both* be givers of the moral law, who thus presumably determine for ourselves what is to be obligatory and what is not, *and* at the same time be governed by it as if we were subjects over whom it stands in the manner of a divine authority?' (Stern 2009:396)

'The idea of a subject, prior to there being a binding law, authoring one and then subjecting itself to it is extremely hard to imagine. It always seems that such a subject could not be imagined doing so unless he were already subject to some sort of law, a law that decreed he ought so to subject himself, making the paradox of this notion of "self-subjection" all the clearer.' (Pippin 2000:192)⁹⁴

Once the notion of self-bindingness is transferred to the theoretical domain, it threatens to undermine theoretical rationality as well. Closely related to this objection is a third: the idea that people's transcendental freedom is a matter of self-bindingness in the theoretical domain, i.e. making judgements, acknowledging claims, having beliefs, may raise the suspicion that this demarcation between human and non-human beings only presents yet another version of transcendental freedom as '*causa sui*'. The original version that was presented by Kant himself, captured in his notion of 'spontaneity', was soon interpreted too by some of his early readers as a form of '*causa sui*'.⁹⁵

If it was correct that Brandom treats theoretical rationality as identical with transcendental freedom, then the criticisms of creating either a Kantian paradox or a *causa sui* would be correct. But the picture of theoretical rationality in Brandom's account is more complicated, I think. Theoretical rationality implies not just *one* but a *double bindingness* to concepts.⁹⁶ It is not simply *identical* to having transcendental freedom. It functions as the *hinge* between both levels of freedom. This would imply that theoretical rationality has a somewhat different meaning depending on which level the notion is used. Brandom does not himself talk explicitly about theoretical rationality as

94 For similar criticisms: see Larmore (2008).

95 See Nuzzo (1994).

96 This is obfuscated by some of Brandom's appreciation of Kant's notion of autonomy as providing the idea of self-bindingness. Although I think that Brandom's account allows for a distinction between two forms of conceptual bindingness, it is not made explicit by Brandom in the way that I will suggest here.

a two-level notion.⁹⁷ For a helpful comment, I will instead turn to Schnädelbach:

‘The predicate ‘irrational’ can be understood in at least two senses: as the contradictory and as the contrary counterpart of ‘rational’. In the one sense, it means the non-rational, the alien-to-the-rational or the rational-less; in the other sense, the irrational, the counterrational, the hostile-to-the-rational. Depending on what counterpart of the rational is intended, the rational itself is differently emphasised as the counterpart of that specific counterpart, and in this way, the broader and narrower meaning of the predicate ‘rational’ becomes clear’ (Schnädelbach 1992b:94).⁹⁸

The difference between both levels may be understood as one between theoretical and instrumental rationality.⁹⁹ The alignment of personhood with rationality – and for their two levels: of transcendental freedom with rationality as opposed to non-rationality (‘das Nichtvernünftige’); of moral-political freedom with rationality as opposed to irrationality (‘das Unvernünftige, Widervernünftige’) – opens up the possibility of making sense of the category of actions for which people are not responsible, but which nevertheless belong to beings that have transcendental freedom.¹⁰⁰ As Schnädelbach himself explains, ‘only *rational* beings may err, act against rules or make mistakes’ (1992b:97, my translation).

My claim is that we will find, corresponding to both levels of theoretical rationality, two different forms of conceptual bindingness. The challenge, however, is to explain how theoretical rationality can be used, as *one and the same* concept, at two different

97 His own rendering of his views on transcendental freedom, in both his systematic and historical writings, make it hard to assess whether this commitment can be ascribed to him. He starts his argument on transcendental freedom with a discussion on autonomy as basically a moral-political notion, as we have seen. But this does not yet mean that he understands rationality as opposed to both non-rationality and irrationality. When next he shifts from autonomy as a moral-political notion towards theoretical self-bindingness as a transcendental notion, this neither indicates a commitment to a two-level notion of rationality. Brandom should nevertheless be understood as being committed to a two-level notion of rationality as will become clear, I hope, from my discussion.

98 My translation. In the original German it says: ‘Das Prädikat ‘irrational’ kann zumindest in zweifachem Sinne verstanden werden: als das kontradiktorische und als das konträre Gegenteil von ‘rational’. Einmal bezeichnet es das Nichtvernünftige, Vernunftfremde oder Vernunftlose, zum anderen das Unvernünftige, Widervernünftige, Vernunftfeindliche. Je nachdem, von welchem Gegenteil des Rationalen die Rede ist, wird das Rationale selbst als das Gegenteil des jeweiligen Gegenteils verschieden akzentuiert, und so treten der weitere und der engere Sinn des Prädikats ‘rational’ deutlich hervor.’

99 Schnädelbach understands rationality at the metaphysical level (‘rational,’ as he calls it) in terms of comprehensibility (*Verständlichkeit*, 1992b:97), which agrees, I think, with how Brandom understands rationality at the metaphysical level.

100 I think, if we bear this distinction in mind, it might explain why the strategy that, for example, Korsgaard, Richard Moran and Isaiah Berlin choose, is bound to fail. In ignoring (or overlooking) this distinction, they seem to have focussed on theoretical rationality as supporting instrumental rationality (moral-political level), in this way overlooking a more fundamental level at which theoretical rationality is constitutive of our freedom, not at a moral-political level, but at a metaphysical level.

levels – as both a contradictory and a contrastive notion: in relation to respectively non-rationality and irrationality – and how, at each level, it agrees with a different kind of conceptual bindingness. The key to answering this challenge is to understand the basic notion of theoretical rationality as constituted by authority *and* responsibility and to understand the difference between the two forms of conceptual bindingness as a difference between *two perspectives* from which theoretical rationality can be analysed: a local and a global perspective.

When Brandom hints at Kant's notion of autonomy, he analyses theoretical rationality from a local perspective, i.e. the attitude that 'concept-mongering creatures' (as he calls human beings) have towards specific concepts, i.e. in specific contexts. This analysis brings out one special form of conceptual bindingness: self-bindingness.¹⁰¹ A simple example may suffice. If someone comes to believe (in Brandom's vocabulary: undertake a commitment) that some animal is a fox, the concept of a fox, as she applies it in this case, does not compel her to say or believe (in the empirical sense of having a mental state with this content) that this animal is a mammal. But it does compel her in the sense that, if she thinks this animal is a fox, she ought not to deny that it is a mammal (or to put it differently: she has a reason to claim it is a mammal).¹⁰² A belief that some animal is a fox is not incompatible with a belief that some other animal is a dog. But the first is incompatible with a belief that the same animal is a dog.¹⁰³ The relevance of the focus on authority is that it allows further analysis of what people are committed to, based on what they say.

Theoretical rationality sustains, next to self-bindingness, yet a second form of bindingness. To understand it, we do not need a different notion or alter the previous notion of theoretical rationality. We just use that same notion of theoretical rationality (*as* consisting of authority and responsibility, *as* inferential) but analyse it from another perspective. Whereas from a *local* perspective, theoretical rationality can be analysed as the attitude conceptual creatures have towards *specific* concepts, from a *global* perspective, it can be analysed as their attitude towards *all* concepts. It implies a different kind of bindingness which is not self-bindingness because conceptual creatures are bound to the

101 My discussion of Brandom's account of self-bindingness (i.e. bindingness from a local perspective) as a form of authority is restricted to what Brandom would call 'content-based authority', which is a form of *intrapersonal* authority, next to 'person-based authority', which is a form of *interpersonal* authority. It relates to Brandom's view on normativity as *social* bindingness, which will be discussed in the next chapter. I refrain from discussing it in this chapter as I think my argument does not require it. Taking authority, in this sense, is equivalent to 'force', in contrast to 'content' (Brandom 2006:53-55) or, in Frege's terminology, to *Urteil*, in contrast to *begrifflicher Inhalt* (Frege [1879] 1977). Also see Geach ([1960] 1972a and ([1965] 1972b).

102 Brandom makes clear that the bindingness of concepts requires that we recognise the distinction between our beliefs in an *empirical* sense and those in a *rational* (or: ideal, logical) sense (1994:507-508). In Brandom's view on inferential rationality, it may be correct to ascribe to some person some specific belief, in a rational sense, even if that person does not have actually have that belief (in an empirical sense).

103 In a slightly different form, this example can be found in Brandom (2009:43-44).

use of concepts as such. While people may choose to endorse some concepts and deny others, they cannot choose *not* to use any concepts at all; they cannot untie themselves from all conceptual norms, and they cannot choose to *deny* all concepts because denying itself is already a conceptual activity.¹⁰⁴ Whereas local bindingness relates to the possibility that people either endorse or deny certain claimables, global bindingness relates to the *impossibility* of either endorsing or denying *all* claimables. Once again: a conceptual creature cannot escape being a concept-user.

The sense in which this absence of choice implies a second form of bindingness, is expressed by the claim that people are situated, in a phrase that Brandom borrows from Sellars, in the ‘space of reasons’. Brandom’s basic notion of theoretical rationality is *inferential*: concepts owe their normative content to the inferential role they have in reasonings. The content of a concept tells us what counts as a reason to use this concept. The dominant strategy in Anglo-American philosophy is to understand concepts as representations of the world that people perceive, believe or judge as true (more on this in section 6). For this strategy, what comes first in the order of explaining knowledge is to explain how beliefs can be true because it is part of this strategy to assume that ‘belief aims at truth’ Underlying this strategy is what we might call a *verificationist* notion of rationality: belief aims at truth. Brandom chooses an alternative, deflationary strategy for which he relies on a tradition within modern philosophy that he traces to Kant, the early Frege and Sellars.¹⁰⁵ They all share the assumption, according to Brandom, that conceptual content comes first in the order of explaining knowledge. What matters for them are not *true beliefs*, but believable or, the term Brandom seems to prefer, *claimables* (or: judgeables, meaning).¹⁰⁶ A crucial move in Brandom’s account is to understand the conceptual content of concepts in terms of their role in reasoning: their *inferential* role, i.e. the role that the conceptual content of a concept plays in an inference. For this reason, concepts are always linked to other concepts. We cannot apply one concept without other concepts:¹⁰⁷

‘one cannot have *any* concepts unless one has *many* concepts. For the content of each concept is articulated by its inferential relation to *other* concepts. Concepts, then, must come in packages’ (Articulating Reasons:15–16)

104 Also, see Brandom (1994:293-297). Richard Moran seems to make a similar point (2001:150).

105 Brandom’s approach belongs to a group of views that start from a deflationary notion of truth and which share the conviction that truth is conceptually fundamental: it cannot be explained in more basic terms (see Armour-Garb and Beall 2005). What is specific about Brandom’s deflationary view is that it understands ‘is true’ as a prosentence forming operator, while for other deflationary views, ‘is true’ is a genuine predicate.

106 The ambiguity of the term ‘belief’ (also see note 99) allows people to easily slip between two senses of the word: its empirical and its rational sense (2000:174). For this, Brandom favours the term ‘claim’ over the term ‘belief’.

107 Behind this idea lies Brandom’s commitment to semantic holism. It also underlies his Kantian account of the self as the product of conceptual activity that ‘produces, sustains, and develops a synthetic unity of apperception: a *self* or *subject*’ (2009:36). I take this account (see 2009, chapter one: ‘Norms, Selves, and Concepts’) as supporting my claim that Brandom too starts from the idea of (a person as) a unity of consciousness.

Conceptual content is normative in the sense that this content itself provides the norm for how it is applied correctly, i.e. what ought to count as a *reason* for specific beliefs, judgements, assertions, claims.¹⁰⁸ People understand the concept ‘that is red’ when they understand, for example, that making this claim is incompatible with claiming ‘that is green’.¹⁰⁹

The metaphor of the ‘space of reasons’ is helpful in contrasting two approaches to understanding conceptual bindingness. The voluntarist approach, as we have seen in chapters 2 and 4, does not necessarily talk of reasons: it understands the space of reasons as a place we enter *after* we have committed ourselves to some choice (or, for that matter: goal, value). The Kantian approach (at least in Brandom’s version) sees people as being already *in* the space of reasons: ‘*all* awareness is a conceptual affair’ (Brandom 2015:111). They cannot escape it.¹¹⁰ The mistake would be to assume that people only enter the space of reasons once they have settled on the goal (value) of their choice and then deliberate on what means will accomplish that goal. It is part of their human condition, however, that people already find themselves in the space of reasons, even before they have reflected on their second-order desires.

‘In a strict sense, all a Kantian subject can do is apply concepts, either theoretically, in judging, or practically, in acting. Discursive, that is to say, concept-mongering creatures, are normative creatures-creatures who live, and move, and have their being in a normative space.’ (2006:56).

People may not be bound to a particular inferential pattern of reasoning unless they commit themselves to it, but at least they are bound to the space of reasons. What it means to be transcendently free is, therefore, paradoxically, to be bound to the space of reasons, not because it entails freedom in an ordinary, voluntarist or instrumental sense, but because bindingness to the space of reasons is a different one than causal-mechanical bindingness. The Kantian paradox exists only within the voluntarist approach. Within the Kantian approach, it dissolves.

108 The idea that meaning is normative and its normativity is irreducible, i.e. that norms are, in Brandom’s much-quoted phrase, ‘all the way down’ (1994:44), is contested for example by Glüer (1999) and Glüer and Pagin (1999), who distinguish between constitutive and prescriptive norms and deny that constitutive norms can be prescriptive. For an interesting discussion of their arguments: Clausen (2004) summarises one of their arguments against the prescriptivity of meaning as follows: ‘if there is something that can guide action or be a reason for it, this means that the action could also take a different course, that one is free either to perform or not perform the action in question. Only if such freedom exists can one be obligated to perform a particular action; in this case one can talk of there being prescriptive norms. But constitutive norms do not guide action’ (2004:113-114). I disagree: constitutive rules can still be prescriptive in this sense. I will return to this issue in the next chapter. For a different defence of Brandom’s notion of prescriptivity: see Peregrin (2014).

109 For verificationists, by contrast, believing ‘that is red’ means having a representation that is true, i.e. corresponds to something that is red.

110 Also, see Turbanti (2017:33) for a similar observation. Admittedly this view is hard to accept for Brandom’s adversaries.

What connects the two forms of bindingness is that bindingness is ultimately a matter of conceptual responsibility because the *original act* of a conceptual creature is not the act of taking conceptual authority, but the act of taking conceptual responsibility: people are born into a state of responsibility.¹¹¹ The difference between the two forms of bindingness arises only by homing in on one or the other perspective. From the local perspective, theoretical rationality is analysed by starting from authority and ending with responsibility. But the conceptual responsibility was already there; people were already in the space of reasons. From the global perspective, the move is therefore reversed: starting from responsibility and ending with authority. It implies that authorising (*self-bindingness*) is only a later stage in becoming a conceptual creature. What this means, or how this works, requires a developmental and social account. I will get to this in the next chapter. For the present chapter, it is important to understand that having transcendental freedom (being in the space of reasons) is possible only by being involved in *practices* of theoretical rationality. This will be discussed in the next section.

5.6. Normative bindingness as practical

Brandom's idea about the normative space of reasons might tempt us into thinking he is committed to the notion of deliberative rationality (see section 4). After all, in his account of rationality, people seem to be highly reflective of their own and other people's commitments and seem to have exquisite skills in articulating and explaining the reasons for what they do and assert. However, Brandom's notion of inferential rationality is more basic: it is practical. The key to understanding how Brandom theorises this option is his *pragmatist turn*, expressed in the slogan: *meaning is use*. This will be explained in the present section. It will show that for Brandom, theoretical rationality (or 'semantics', as Brandom usually calls it), as a basis for transcendental freedom, is really a form of practical rationality, not in its narrow, standard sense of rationality as instrumental, i.e. as the 'calculation' of what is the right means to some goal, but in a much broader sense.

Conceptions of deliberative rationality that are understood in terms of 'making inferences' are sometimes illustrated with one specific type of 'inference' one might make: the *modus ponens*. It provides a way of inferring the truth of a claim from the truth of two other claims. Its formal structure can be presented like this:

- (A) $x \rightarrow y$
- (B) x
- (Z) $\vdash y$

111 As Brandom (1994:xii) puts it: 'the characteristic *authority* on which the role of assertions in communication depends is intelligible only against the background of a correlative *responsibility* to vindicate one's entitlement to the commitments such speech acts express'.

The reason for its popularity in certain philosophical literature seems to be that it provides a prominent yet simple example of a normative conception of rationality, and so, if it can be shown that the *modus ponens* is defective, then this normative conception, too, is defective. The *modus ponens* has therefore raised much discussion in contemporary discussion. Perhaps its first critical treatment has been Lewis Carroll's 'What the Tortoise said to Achilles' (1895). Even if it does not mention *modus ponens* as the focus of its discussions, at least it has been the point of reference for many subsequent discussions because it introduces the problem of the regress of rules: the Regress Objection.¹¹² It shows an important weakness of the deliberative view on rationality (if by this we understand the view that identifies 'being rational' with 'following rules', in this case: inferential rules) by introducing an argument that seemingly undermines the validity of the *modus ponens* as a rule of inference. What was the Tortoise saying to Achilles? Here is the pattern of reasoning that starts their discussion (and which I take the liberty of rephrasing as a *modus ponens*):

- (A) If things are equal to the same, they are equal to each other.
- (B) The two sides of this triangle are things that are equal to the same.
- (Z) Therefore: The two sides of this triangle are equal to each other.¹¹³

Achilles claims that the conclusion follows logically from the two premises: any who accepts the truth of (A) and (B) *must* accept the truth of (Z). This is contested by the Tortoise who invites Achilles to agree with him that an extra premise is needed:

- (C) If (A) and (B) are true, (Z) must be true.

Achilles agrees and believes adding this premise will finally convince the Tortoise that anyone accepting (A), (B) and (C) *must* also accept the truth of (Z). The Tortoise does not give in. Once again, the Tortoise claims the possibility that some might deny the truth of premise (A), (B) and (C) unless a further premise is added:

- (D) If (A) and (B) and (C) are true, (Z) must be true.

Achilles agrees. By now, the point must be clear to the reader: the moral of Carroll's brief story is that Achilles may have no hope of ever escaping this argument because the Tortoise will continue adding new steps without the conclusion ever coming nearer.

Carroll's story is about the problem of justifying the conclusion one draws from a certain pattern of reasoning. Once we generalise this problem to the body of all

112 Better known as the 'Frege point' (Geach ([1965] 1972b:255) or the 'Frege-Geach challenge' (e.g. Finlayson 2005).

113 "That beautiful First Proposition of Euclid' the Tortoise murmured dreamily' (Carroll 1895:278).

reasoning, i.e. to the validity of what we take to be our knowledge, a more fundamental problem arises: how do we justify our knowledge? This is not the problem of justifying just one case of reasoning, as in the dialogue between Achilles and the Tortoise, but of all our knowledge. For example, Hans Albert (1968/ 1991) has argued that once a justification is demanded for *all* of our knowledge, the knowledge that is invoked for this justification is itself in need of a justification.¹¹⁴ This problem may be solved by taking recourse to either of three alternatives: (1) entering into an infinite regress, (2) getting caught in a logical circle or (3) forcing a termination at some point.¹¹⁵ As Albert suggests, all three seem unacceptable and therefore present a trilemma, what he appropriately calls the ‘Münchhausen-trilemma’. The third option may be said to be illustrated by Frankfurt’s account of transcendental freedom: his idea of a second-order desire seems to elicit even further orders of desire.¹¹⁶

As we have seen (previous section), Brandom’s account starts with Kant’s lesson: freedom consists of how we human beings follow rules. We follow them according to our grasp of norms. But Brandom does not go all the way with Kant. For Kant, Brandom explains, ‘rules are not simply one form among others that the normative might assume. Rules are the form of the norm as such’ (1994:19–20). Action, which is understood as applying concepts, should be judged according to explicit rules because they tell us the correct use of a concept and, therefore, whether the action was correct. At this point, he disagrees. He rejects the position of which Kant is a representative and which Brandom calls *regulism*.¹¹⁷ Explicit rules are not exhaustive of what norms are because they themselves require explanation. Therefore, you always need new rules to explain how the previous rules are applied correctly in an endless regress.¹¹⁸

Rules are not the only form of normativity. Brandom believes that explicit rules presuppose a form of normativity that precedes the normativity of explicit rules: ‘norms that are explicitly expressed in the form of rules, which determine what is correct according to them by *saying* or describing what is correct, must be understood as only one form that norms can take. That form is intelligible only against a background that includes norms that are *implicit* in what is *done*, rather than *explicit* in what is *said*’ (1994:30). People’s grasp of a situation is not evinced by their ability to *tell* what rules pertain to that situation but by their ability to act correctly in that situation. In Brandom’s account, understanding acting according to rules is replaced by an understanding of acting according to mastery ‘where the performer is being guided by something, but not

114 Albert (1968/ 1991, first chapter).

115 Infinite regress (*infiniter Regreß*) is one of three alternatives that Albert (1968/ 1991:15) distinguishes for solving the problem of finding a foundation for knowledge.

116 Suggested by Kornblith (2012:74–78), although Frankfurt seems to deny this (1971). Also, see Frankfurt (1987).

117 Following Dewey, Brandom also talks about ‘intellectualism’ or ‘platonism’ (2008:40) in contrast to his own position (pragmatism).

118 Brandom also mentions Wittgenstein’s regress-of-rules argument and Sellars’s regress-of-interpretations argument (1994:20-26).

by something explicit and articulate' (1994:65). He then makes a move from practices that display a person's grasp of some situation (e.g. knowing the difference between a dead and a living parrot) towards the attitudes that underlie these practices. These attitudes are normative. Normative attitudes display a 'sensitivity to norms that [Kant] talked about in terms of conceptions of rules' (Brandom 1994:45). Beings that have a normative attitude are directed towards the world, move around in the world, deal with the world always in a sense that they treat whatever they encounter from a normative perspective, taking the world not just as it is, but as it ought (not) to be.

In the search for an alternative view, Brandom takes over two lessons from Hegel and Wittgenstein: we should understand norms ultimately in terms of practices, and we should understand them as social practices: 'norms in terms of practices rather than exclusively in terms of rules, and the recognition of the social character of such norms' (MIE:55). But what complicates understanding the normative, socio-pragmatic view that Brandom expounds in the first chapter of his *Making It Explicit* is that he chooses to explain it as part of a polemic with various forms of naturalism. For these, he uses the term 'regularism' because they agree that human behaviour should be understood in terms of regularities, dictated by laws of nature, rather than rules people have chosen themselves.¹¹⁹ In this polemic, he weaves in two explanatory strategies he borrows from others. The first strategy he owes to Haugeland, who suggests that we imagine 'a community of versatile and interactive creatures' (Haugeland 1982:15) as the starting point for finding out what it means that people are social beings. But there is a twist here. If rationality is, in its basic form, practical, then it would seem to imply that beings might be rational without being linguistic. Brandom calls them 'proto-hominids': 'what proto-hominids could do before they could talk is to take or treat each other's performances as correct or incorrect by practically sanctioning them, e.g. by beating each other with sticks as punishment' (1997:196).¹²⁰ Brandom then discusses various views of how their practices might alternately be interpreted as either normative or regular. Finally, he takes recourse to one other strategy he takes over from Dennett and introduces the distinction between simple, intentional systems and interpreting intentional systems (or: original intentionality).¹²¹ It allows him to explain how the practices of a community are being interpreted as normative by an outsider: the Interpreter.

My interest here is not with the details of how Brandom works out both strategies, but with his use of these strategies: he seems to use them to understand normative attitudes as rooted in social practices and to understand social practices as being interpreted as

119 In the literature on the concept of rule, especially when related to Wittgenstein's views, these two forms of acting according to rules are usually distinguished as rule-conforming versus rule-following: see cf. Bloor (1997); Gerrans (2005).

120 This phrase is not used in *Making It Explicit*. It is introduced only later, in 'Replies' (1997:196).

121 See Dennett (1971), (1987, esp. chapter 3: 'Three Kinds of Intentional Psychology'), (1990), (1990). For Dennett, the intentional stance only indicates a strategy of making sense of intentional states without the need to assume that these intentional states (beliefs) really exist in the heads of, for example, people.

normative by an Interpreter who does not participate in these practices. This seems to be the basic argument of Brandom's normative, socio-pragmatic account:

1. *Meaning is explained in terms of use*: norms explicit in what is said (here: norms as *rules*) are explained as norms implicit in what is done (here: norms as *correct performances/ practices/ actions*). Norms implicit in what is done can be further explained in terms of having a normative practical attitude, i.e. the ability for correct performances.
2. *Correct or incorrect performances are explained in terms of the sanctioning practices of a community*: the members (proto-hominids) of a community attribute to each other normative practical attitudes, and they treat each other's performances as correct or incorrect by sanctioning them – i.e. treating a performance as correct by rewarding its performer and treating it as incorrect by punishing its performer.
3. *The sanctioning practices of a community (including the sanctioning practices) are explained in terms of the attributions of an 'interpreter'*: an outsider to the community (a hominid?) interprets the practices of this community, including its sanctioning practices, as exhibiting normative, rather than merely regular, behaviour by attributing to its members a normative practical attitude.

Brandom's account has been criticised by various theorists whose remarks abound with observations that his normative pragmatism is also vulnerable to the Regress Objection.¹²² They usually refer to the first chapter of *Making It Explicit* (and to a lesser extent to its final chapter), in which Brandom chooses to introduce his normative-pragmatic argument as a response to what poses a problem for the traditional notion of deliberative rationality ('regulism', as he calls it): the Regress Objection. A different, and I think *better*, explanation of the account Brandom is committed to, and of how it

122 Turner (1998/ 2002a); Glüer (1999 and 2002); Wikforss (2001); Hattiangadi (2003; 2006 and 2007); Grönert (2006); Glüer and Wikforss (2009); Kornblith (2012). All of them mention Brandom and others as representatives of the 'meaning is normative' claim. Some of them (in some of their writings) also explicitly focus on the details of Brandom's argument. Oddly Kornblith (2012) treats Brandom as a representative of theorists with reflective ('higher-order') views of rationality. This misrecognises Brandom's pragmatism on normativity. Unfortunately, Kornblith does not discuss any details of Brandom's argument. For various defences of Brandom's account: Clausen (2004); Peregrin (2014); Turbanti (2017). It is not entirely clear how Brandom's critics understand the notion of regress. It might be a *simple* notion of regress, one that states that any account turns into a regress if it provides an explanation of A by reducing A to B, and an explanation of B by reducing it to C if B cannot function as a regress-stopper (because B requires yet another explanation), and so on. Understood this way, Brandom's account may be said to run into a regress. I nevertheless think that Brandom's critics have misunderstood what I present here as claims (1), (2) and (3). The misunderstanding can be traced to how they treat Brandom's successive explanations (explaining explicit rules as practices, practices as attitudes, attitudes as results of sanctions, sanctions as ascriptions), which is to treat them as ever so many reductions (reducing explicit rules to practices, practices to attitudes, et cetera) and therefore as belonging all to the same order, as if there was no difference in kinds of explanations. In this section, I try to show not all claims belong to the same order of explanation.

escapes the Regress Objection, can be found, I think, in his *Between Saying and Doing* (2008a). Here he offers ‘a new way of thinking about language’ (2008a:1), one that analyses language in terms of the relations between what is said and the activity of saying it or, differently put, between meaning and use, between vocabularies and practices-or-abilities.¹²³ (In *Between Saying and Doing* Brandom frequently uses also this second terminological pair. I will also use both.) In Brandom’s technical vocabulary, these relations are called Meaning Use Relations (MUR). Here Brandom does not mention the problem of rule-following. But as the problem of rule-following arises from the identification of rule-following with meaning that is explicit in language, i.e. in what is said or thought (regulism, deliberative rationality), I think the new view he expounds in *Between Saying and Doing* can also be read as an answer to the Regress Objection.

Vocabularies and practices-or-abilities have a *reductive* relation: vocabularies can be understood as reducible to practices-or-abilities.¹²⁴ This is how claim (1) should be understood. The relation between vocabularies and practices-or-abilities is not just reductive (i.e. vocabularies as reducible to practices) but also *expressive* (i.e. practices as expressible in vocabularies). The ascriptions the Interpreter provides – as in claim (3) – illustrate the expressive relation.

Understanding the expressive relation requires a better understanding of Brandom’s pragmatism. His point in making a distinction between vocabularies and practices is precisely to move beyond vocabularies, i.e. beyond what is explicit in what is said (or thought), and therefore to move beyond explanations which, too, belong to the domain of vocabularies (meaning, the explicit, the said). The reductive relation is therefore not about reducing one explanation to another, but about reducing the domain to which all kinds of explanations belong, to a different domain that, strictly speaking, offers no explanations: practices-or-abilities. These are just what they are: performances. They do not talk and therefore do not provide explanations in the standard sense: they do not belong to the same order as do explanations that consist in what is said. Obviously, reducing an explanation (and with it all that belongs to the domain of meaning) to what is not an explanation implies a puzzling claim about the relation between meaning and use. If Brandom proposes to reduce meaning to use, he must, in making this proposal, somehow articulate this strategic move in explanations, in what is said, in *meaning*. So, it raises a question about the status of the kind of vocabulary (the type of meaning)

123 For his use of the phrase ‘practices-or-abilities’: Brandom (2008a:9, note 8).

124 Although much of Brandom’s writings focus on the coexistence of vocabularies and practices-or-abilities, the reductive relation suggests the possibility of having rationality without having language. Brandom himself seems to acknowledge this by introducing ‘proto-hominids’: ‘what proto-hominids could do before they could talk is to take or treat each other’s performances as correct or incorrect by practically sanctioning them, e.g. by beating each other with sticks as punishment’ (1997:196). But from his remarks (see 1997:196, note 5), it also becomes clear that their rationality should perhaps be understood as a kind of proto-rationality. A discussion on the meaning of this would focus on the relevance of having language to having rationality. However relevant that discussion is, it is beyond the purposes of my present argument.

Brandom uses to express that idea.¹²⁵ Brandom's response is that vocabularies can be used to express the norms that are *implicit* in practices: 'Talk of practices-or-abilities has a definite sense only insofar as it is relativized to the vocabulary in which those practices-or-abilities are specified' (2008a:9).¹²⁶ This places vocabularies in an expressive relation towards practices-or-abilities. In *Making It Explicit*, we already encountered the idea that language enables us to make explicit what is implicit in practice. In *Between Saying and Doing*, the expressive power of a vocabulary is indicated by his technical term 'VP-sufficiency', in which 'VP' stands for 'vocabulary practice' and which is 'the relation that holds between a vocabulary and a set of practices-or-abilities when that vocabulary is sufficient to specify those practices-or-abilities' (2008a:10). The reverse, *reductive* relation is indicated by Brandom's technical term 'PV-sufficiency', in which the terms 'vocabulary' and 'practice' are now reversed and which 'obtains when engaging in a specified set of practices or exercising a specified set of abilities [...] is sufficient for someone to count as *deploying* a specified vocabulary' (2008a:9).

Some of Brandom's critics have understood claim (3) as implying a reductive relation: the normativity of sanctioning (communal) practices can be understood as reducible to the attributions of an interpreter. The previous discussion of Brandom's account on meaning-use relations, however concise, already indicates the incorrectness of this interpretation of claim (3). Practices come first in the order of explanation, even if explicating practices requires taking recourse to vocabularies that possibly come later in the order of explanation. Brandom suggests that the notions of PV- and VP-sufficiency allow a more complex understanding of the relation between practices and vocabularies. Some set of practices are needed to count as deploying a certain vocabulary we will call: V. But vocabulary V is not necessarily identical to the vocabulary that explicates the practices underlying vocabulary V. The second vocabulary is strictly a different one. It functions as what Brandom calls a '*pragmatic metavocabulary*' for V because it 'allows one to *say* what one must *do* in order to count as *saying* the things expressed by vocabulary V' (2008:10). Consequently, we should understand the relation between the practices of proto-hominids and the ascriptions of the interpreter not as a reductive but as an expressive one. The ascriptions by the interpreter are expressions of what proto-hominids do, while the practices performed by proto-hominids are not supposed to be reducible to the ascriptions of the interpreter. And so, Brandom's critics are wrong in assuming that the shift from practices to ascriptions is yet another reductive explanation Brandom offers.

The two relations between meaning and use – *reductive* and *expressive* – seem to force a choice between either one of them as more basic, i.e. as coming first in the order

125 However puzzling Brandom's proposal to reduce meaning and use, in this sense, Brandom's critics do not seem to be alarmed about it.

126 He understands this move as part of his project to extend the project of analytic philosophy: Brandom (2008a, chapter 1).

of explanation. And the drawback of not choosing is to remain caught in a logical circle. How may we understand practices as coming first in the order of explanation without giving up on the expressive relation? Brandom's account can escape, I think, both the infinite regress and the logical circle.¹²⁷

First of all, it is important to keep in mind that whatever terminology Brandom uses to explain meaning-use relations, normativity remains central to them. The reduction of vocabularies to practices, as we saw in claim (1), is not a reduction of something that is normative to something that is *not* normative. So, in this reduction, normativity at least remains preserved. Whatever it is that is reduced or somehow 'gets lost' in the process of reduction is something different. But what? The reductive loss can be traced to a change of normativity itself. That Brandom does indeed has this kind of change in mind, appears from his choice of using different terms to indicate this normativity. As we have seen before, the normativity that is explicit in what is said (vocabularies, meaning), he calls 'rules' whereas the normativity that is implicit in saying (practices, use) he calls 'norms'. Of course, making this terminological distinction is not enough. It yet needs to be explained what exactly the difference is between 'rules' and 'norms'. Roughly it is a difference between logical and non-logical normativity. Vocabularies owe their expressive power to logical vocabulary.¹²⁸ What gets lost in the reduction of vocabularies to practices is *logical* vocabulary.

Some of Brandom's critics would reject the pragmatic reduction implicit in claim (1). Underlying their rejection is a tacit assumption that reasoning requires logical normativity, i.e. that reasoning in principle requires a distinction between form and content: between logical operators (form) and that which can be considered apart from logical operators (content).¹²⁹ But they fail to notice that the normativity implicit in practices differs at this point from normativity explicit in vocabularies. Even more so, they fail to notice that, in Brandom's account, it is because of this implicit, i.e. non-logical normativity, that practices come first in the order of explanation: rational beings need to have mastered non-logical normativity in what they do before they can use logical vocabulary in what they say.¹³⁰

127 As Dennett's account is not at the centre of my argument, I will not explore whether this is too, in the end, able to escape the logical circle.

128 This expressive power allows, for example, the Interpreter to understand the reduction of vocabularies to practices. It may even be claimed that it allows the Interpreter to better understand the proto-hominids' practices than they do themselves.

129 See Glüer and Wikforss when they conclude that the implicitness introduces some kind of regress again: 'this time the regress is one of implicit norms, but that does not make it any less bottomless' (2009:63). Underlying their analysis, I think, is still a *formal* conception of what norms are, i.e. as contents that can be distinguished from their (logical) form. If this is correct, then they are right to conclude that a new regress starts. But this is not how Brandom understands the implicitness of norms.

130 I also believe that they misunderstood Brandom's notion of reduction in yet a different sense, i.e. they conflate normativity and correctness, i.e. what is *treated as correct* and what *is correct*. Brandom's account of normativity should be distinguished from his account of correctness. What is at stake in this discussion is how to understand that practices of taking-as-correct can be a source of objective knowledge. As my argument is not primarily concerned with this aspect of Brandom's account, I will not further explore this misconception. However, see Clausen (2004).

Below I will explore how Brandom understands normativity implicit in practices and how this explains the priority of practices over vocabularies. The challenge is to explain how Brandom understands *non-logical* inferences and how these precede logical inferences.

Next to PV-sufficiency, Brandom distinguishes PV-*necessity*, which ‘obtains when one cannot deploy a certain vocabulary without engaging in the specified practice, or exercising the specified ability’ (2008a:28). It allows us to ask what *specific* practices-or-abilities are necessary for deploying *specific* vocabularies, such as observational, logical or normative ones. Brandom is interested in what set of practices-or-abilities is necessary for deploying *any* vocabulary: ‘Are there any practical abilities that are *universally* PV-necessary?’ (2008a:41). The set of practices-or-abilities that is necessary for any vocabulary is what Brandom calls an autonomous discursive practice. Here I will discuss only one subset: the *inferential* practices-or-abilities. People can make inferences without using logical vocabulary. These kinds of inferences Brandom calls ‘material inference’. But what does it mean to make material inferences? I will focus on two features (that we will describe using a *modal* metavocabulary): *modality* and *material incompatibility*.

Making material inferences involves that a person can distinguish at least some inferences as materially good, i.e. she is able to act in such a way that from her actions, it is clear she understands what follows from what. Brandom himself provides the example of a lioness:

‘One grasps the claim “the lioness is hungry” only insofar as one takes it to have various consequences (which *would* be true if it *were* true) and to rule out some others (which *would not* be true if it *were* true). And it is not intelligible that one should endorse as materially good an inference involving it, such as the inference from “the lioness is hungry” to “nearby prey animals visible to and accessible by the lioness are in danger of being eaten,” but be disposed to make no distinction at all between collateral premises that would, and those that would not, if true inform the inference. One must make some distinction such as that the inference would still go through if the lioness were standing two inches to the east of her actual position, the day happened to be a Tuesday, or a small tree ten miles away cast its shadow over a beetle, but not if she were shot with a tranquilizing dart, the temperature instantly plummeted 300 degrees, or a plane crashed, crushing her’ (2008a:105).

Part of what it means to make material inferences is therefore that a person not only treats some material inference as good, but also treats it as including certain collateral claims, i.e. a certain range of counterfactual robustness: ‘One has not grasped the concept cat unless one knows that it would still be possible for the cat to be on the mat if the lighting had been slightly different, but not if all life on earth had been extinguished by an asteroid-strike’ (2008a:97).

Another feature of making material inferences is the ability to recognise material incompatibility. It can be expressed by using alethic modality as a metavocabulary to express what people do, i.e. recognise material incompatibility. This metavocabulary can also be used to restate the *modus ponens* in modal terms, without using a conditional:

“ $x \rightarrow y$ ” is equivalent to “It is not possible that x and *not-y*”¹³¹

And therefore:

It is not possible that x and *not-y*.

x

$\vdash y$

An important of this modal metavocabulary is the negation: ‘negation lets one make explicit, in the form of claims – something that can be said and (so) thought – a relation that otherwise remained implicit in what one practically did, namely treat two claims as materially incompatible’ (2008a:48). Now people can recognise material incompatibility without needing to use logical vocabulary, not even the negation. As Cogburn summarises Brandom’s point:

‘speakers of a language without a conditional or negation still have ways of accepting or rejecting material inferences. If I say that something is water, and then you say that that thing is dry, I remonstrate with you. If you say it is wet (and I’m teaching you the meaning of the word, for example), I will praise you. If you say of the same thing that it is both dry and wet, I remonstrate with you’ (2010:168).

What is expressed in a *metavocabulary* using logical operators can be done practically, i.e. *without using* a vocabulary with logical operators. Because this is possible, practice precedes vocabulary.

If we accept that the ability to make material inferences, which in a sense is an ability that does not yet rely on a formal distinction between form and content, *precedes* the ability of formal inferences, then it will not be necessary that Achilles is able to make explicit what makes inferences valid, because the validity of inferences in a way is a feature that is part of the practice of treating something as an inference, that than something that can be made fully explicit. To put it differently and somewhat simpler: the solution to the problem that the Tortoise poses to Achilles, the Regress Objection, is already suggested in the opening line of Carroll’s article: Achilles is seated on the Tortoise. As soon as their conversation starts, it becomes clear what this means. Achilles has solved the paradox of the

¹³¹ Also, see Brandom (2008a:109).

fast runner trying to catch up with the slow runner, not by theorising about running, but plainly and simply by doing what should not be theorised: by running (or just walking: *solvitur ambulando*). The problem of the validity of the modus ponens can be solved in a similar way. Achilles might confront the Tortoise with a simple puzzle, picking him up and holding above a hot pan of soup, giving him a choice: 'If you give the right answer, I will put you on the ground again. If you give the wrong answer, I will let you fall'. As it will turn out, or not, the Tortoise will show his ability to grasp the concept of a modus ponens without any further explanation. And this shows that people (and perhaps tortoises as well) can grasp what they cannot always express.

In this chapter, I have outlined Brandom's normative pragmatism. This paves the way for replacing the mentalist framework with a Brandomian-pragmatist framework. In the next chapter, I will continue my explanation of Brandom's normative pragmatism and show how it can make sense of the veil of prejudice without taking recourse to conscious/unconscious vocabulary. Brandom's normative pragmatism has also provided us with a kind of rationality that is different from what is often assumed in the liberal account of autonomy: not *deliberative* rationality, not *instrumental* rationality, but a more basic, *inferential* rationality. It is not what largely belongs to the better educated and more articulate amongst us. It is what most of us, human beings, have. In the final chapter, I will explore how this notion may sustain a moral-political notion of rationality.

5.7. The Cognitive Suicide Argument

The Manipulation Argument was introduced in the previous chapter as undermining the standard (i.e. *liberal*) notion of freedom and responsibility. In the preceding sections, I have tried to show that Brandom's account of freedom and responsibility is not vulnerable to the Manipulation Argument. In this concluding section, I will go one step further. I will show how his account provides an argument against the *validity* of the Manipulation Argument. It will be called the Cognitive Suicide Argument.¹³² The Manipulation Argument is designed to convince us of the incompatibility of determinism and freedom. But it is itself built upon an incompatibility that it cannot solve. It presupposes not only a claim about the causal chain of the world, the Determinism Thesis but also the claim that people are *knowers*. The problem, however, is that these two claims are incompatible. Either determinism is true (and people cannot be knowers), or people are indeed knowers (and determinism is false).

132 For the phrase 'cognitive suicide': see Hanna (2008). For a similar argument: see Lance and White (2007), who also draw on Brandom's insights.

A commitment to the Determinism Thesis is therefore committing cognitive suicide.¹³³

I start by briefly outlining what I call the Modality Argument. As we have seen in the previous section, Brandom assumes that modality is a feature of rational thinking: people have the ability of thinking in terms of possibilities. If determinists grant at least this much to Brandom and agree with him on this view, then what is at stake here is the question of whether modality of thought also indicates that modality is a feature of the world itself.¹³⁴ Denying this, as I assume that determinists would, makes it hard to understand how people can be knowers in a deterministic world. The problem would not just be that the coincidence between people's action and their will (or, within the Brandomian-pragmatist frame: between their action and their reasons) is 'only a happy chance', as Frankfurt puts it. The problem would be to understand at all how people can have rationality (either the instrumental rationality that Frankfurt assumes or the broader rationality that Brandom argues for). If people cannot tell the difference between an action that is the result of their will, and an action that is the result of processes that are completely unknown to them, then how could they ever acquire knowledge?

The Manipulation Argument is undermined by the Modality Argument. But the counterargument might be that modality does not exist as a normative force in either the physical world or the mind. In this view, not even the inner citadel is safe: the mind itself, including the notion of modality, is determined. This idea is not new. It was accepted already by various early modern philosophers (section 2). And yet, they also recognised that people could acquire knowledge. Some philosophers, such as Geulincx and Leibniz, assumed that the workings of the body (movements, actions) and the workings of the mind (will, emotions) not only are determined but also run synchronously. To make this assumption plausible, Geulincx introduced his well-known metaphor of two clocks agreeing precisely with each other.¹³⁵ This still leaves the metaphysical question of how

133 Several prominent philosophers have argued against determinism. For example, Dennett has argued against determinism in many publications, partly against the conception of freedom as 'could have done otherwise', often called the issue of 'alternative possibilities' (1984b), but also against the Manipulation Argument (1973 and 1984). One of his main arguments is directed against the hidden assumptions in many examples of manipulation or what Dennett calls 'intuition pumps'. Other prominent critics are Bhaskar (1975, especially chapter 2) and Habermas (2005 and 2007), who is particularly critical about it, remarking that the 'ontologization of natural scientific knowledge into a naturalistic worldview reduced to "hard facts" is not science but bad metaphysics' (Habermas 2005/2008:207). For conceptual difficulties with the notion 'determinism', see, for example, Dowe (2002). For an exploration of determinism in relation to theories of modern physics: see Earman (2004).

134 Price (2008) agrees with Brandom's view on the modality of thinking but has suggested that modality is no feature of the world itself. Brandom responded: 'I don't understand what 'a nonmodal world' could conceivably mean. That counterfactual – what if we inhabited a nonmodal world – makes no sense if the expressive resources recruited by modal vocabulary are really part of the very meanings of the words that we use to describe the world' (2008a:145). Elsewhere Brandom also writes: 'A world without modal facts would be an indeterminate world: a world without objects in the Aristotelian sense, and without properties in the sense that admits a determinate-determinable structure' (2015:203). For a suggestion similar to Price's: see De Caro and Voltolini (2010).

135 Geulincx (1665/ 1968:212, *Annotata ad Ethicam*, § 19). The clock as a machine model for the necessary workings of the mind was used more often in the early modern period: McReynolds (1980).

to explain the coincidence between the body and the mind. Whereas early modern philosophers still invoked God, hard determinists, of course, rely on the laws of nature. It is unclear, however, from the determinism debate whether determinism (of the physical and the psychological realm) also extends to the realm of rationality. To put it more simply: is some specific reasoning (only) a psychological process that obeys the laws of nature, or is it also, in some sense, normative? Hard determinists such as Pereboom do not seem to deal with this question even if their Manipulation Arguments tacitly assumes that a ‘reasoning pattern’ can be part of a manipulated process. One who does provide an explanation for it is Carruthers. For that reason, I will, in this section, explore his views on this issue. I will argue that these also undermine the Manipulation Argument.

I will also take up an issue that I have not yet resolved: how to explain that a person may be mistaken? In section three, I suggested the option that some actions might be described as transcendentally free because they can be ascribed to a person and that this person might yet be non-responsible because that person was only in error for reasons she was not to blame. This option presupposes a two-level notion of freedom. If it is possible to make this option theoretically plausible, this will undermine the Manipulation Argument: it would sustain our ‘intuition’ that Plum, as presented in Pereboom’s thought experiment, is manipulated, but this would not yet imply that Plum has no transcendental freedom. In short: the case of Plum, instead of vindicating hard determinism, would by itself be inconclusive. In section five, I have argued that Brandom is committed to a two-level notion of rationality. He understands rationality as non-mechanical bindingness that is both normative and social. This offers a strategy for demarcating some objects from other objects for being transcendentally free, i.e. they are really subjects (persons) rather than objects. But it does not yet explain how a person may be mistaken. How may a person be able to be wrong about treating something as correct? As it will turn out, making mistakes is central to Brandom’s account of rationality: ‘We are discursively born into a state of sin, and, for all our conscientious efforts, are by and large doomed to live in such a state’ (2008a:191-192).

In the previous chapter (section 4), we have seen how Carruthers interprets manipulation experiments in a way that undermines any putative *introspective* claim to having knowledge. But we have not yet seen how Carruthers argues that people may nevertheless, *without introspection*, have knowledge. The dual-process theory of mental processes, a third-person perspective on human cognition, allows for a theory of human cognition from a first-person perspective: reliabilism. As an epistemological view, it can be understood as a response to an older epistemological view that understands knowledge in terms of *justified, true belief*. According to this traditional view, people may claim not only to have *true* beliefs but also to have *justified* true beliefs because they had introspective access to whatever justified those true beliefs. As this view lost ground, due in large part to Gettier’s influential article ‘Is Justified True Belief Knowledge?’ (1963),

an alternative view was provided by reliabilism which replaces ‘reasons’ with ‘reliable process’ of acquiring knowledge. As Carruthers himself explains:

‘knowledge is true belief that is caused by a reliable process. A reliable process is one that generally issues in true beliefs, and that also serves, in the particular case in hand, to discriminate reliably truth from relevant falsehood’ (1992:73).

Carruthers believes that reliabilism also sustains knowledge about more abstract matters because it can explain, as he claims:

‘our knowledge of laws of nature, if the processes of induction and inference to the best explanation employed by scientists are also generally reliable’ (id.).

Reliabilist epistemology, therefore, replaces the concept of having good reasons for belief with the concept of reliability of processes of belief acquisition. It does not have to assume a ‘space of reasons’. Having knowledge is about having *reliable* (rather than *justified*) true beliefs.

It might seem that Brandom disagrees with reliabilists on exactly *which* ability constitutes the basis for acquiring knowledge: reliabilists acknowledge only people’s observational capacity – what Brandom would call a *non-inferential reporting capacity* –, whereas Brandom picks out the inferential capacity. His account of having knowledge does not, however, depend entirely on the exercise of inferential rationality. Brandom would deny the possibility of an ‘inner citadel’ if only because he denies that pure deliberation, without observation, is possible: ‘Purely theoretical concepts do not form an *autonomous* language game, a game one could play though one played no other’ (2002:366). In fact, he subscribes to what he calls the ‘founding insight of reliabilism’ that ‘true beliefs can, at least in some cases amount to genuine knowledge even where the justification is not met (in the sense that the candidate knower is unable to produce suitable justifications), provided the beliefs resulted from the exercise of capacities that are *reliable* producers of true beliefs in the circumstances in which they were in fact exercised’ (2000:97). Brandom’s account acknowledges, therefore, two abilities that *together* constitute the formation of knowledge: next to an inferential capacity also a non-inferential reporting capacity. And he criticises reliabilists for not acknowledging the inferential capacity (in my phrase: they commit cognitive suicide).¹³⁶

Brandom’s account does not yet provide a decisive argument against *mechanical* necessity because the distinction between the two capacities is not enough to oppose

136 For Brandom’s discussion of this point: see Brandom (1998).

Carruthers' account.¹³⁷ Carruthers would agree with Brandom that observational and inferential capacities are both needed to understand the process of acquiring knowledge. It is just that Carruthers *denies* that the normative domain, whether we call it inferential rationality, the 'space of reasons' or the 'sea of normative beliefs', is somehow excluded from the mechanical laws of nature.¹³⁸ Carruthers' counter-strategy has been to develop a specific variety of reliabilism that not only explains true beliefs in terms of reliability but also shows, now paraphrasing Brandom, how correct reasonings result from the exercise of mechanisms that are *reliable* producers of true reasoning in the circumstances in which they were in fact exercised. The first step in this strategy is to explain concepts in terms of innate structures, which, drawing on Fodor's theory of modularity (1983), he calls 'conceptual modules'. Conceptual structures (reasoning) are not normative but 'brute-causal'.¹³⁹ Later he integrated the idea of modules in his views on the mind as consisting of System 1 and System 2.

'On my account it is *beliefs about* norms, rather than norms themselves, that do the explanatory work. Representations involving modal concepts like MUST or MUSTN'T are stored in a special-purpose belief-box, attached to a reasoning system that is continually on the lookout for circumstances that might lead to a match with the non-normative content of those representations. When one is found, the system generates an intrinsic motivation towards performing the action described by that content (in the case of 'must'), or against acting (in the case of 'must not'). Whether the represented norms are true or correct is quite another matter. And the resulting account is fully consistent with a naturalistic, purely causal, account of thinking and believing' (2006:262).¹⁴⁰

137 The argument Brandom develops against reliabilism (2000, chapter 3) is really an argument against empiricism (2002: Chapter 12). Carruthers too opposes empiricism, but, very differently from Brandom, has developed a view, called evolutionary nativism, that remains firmly embedded within a naturalist programme by trying to understand reliabilism as an evolutionary feature not just our observational capacities, but of our cognitive capacities as well (1992:77): 'Innate beliefs will count as known provided that the process through which they come to be innate is a reliable one (provided, that is, that the process tends to generate beliefs that are true)'.

138 Carruthers mentions as representatives of the normative view on cognition ('spontaneity', 'space of reasons', 'norms of reasoning'): McDowell and Brandom (both: Carruthers 2013b:234; 2014 and 2015:10), Korsgaard (Carruthers 2015:10) and Davidson and Dennett (both: Carruthers 2006:262; McDowell is mentioned here again).

139 Also, see Cowie's discussion on Fodor (1999).

140 I believe that this view makes Carruthers' account vulnerable to the Regress Argument. If 'whether the represented norms are true or correct' is in fact 'another matter', this still leaves the question of what it means that some represented norms are true and how this can be assessed. It requires the use of a meta-vocabulary to explain this. And this requires an explanation of how one is able to use a meta-vocabulary. So, the regress starts into a meta-meta-vocabulary. For making a similar point against Carruthers (2009): see Anderson and Perlis (2009).

The explanation of reasoning in terms of conceptual modules (reasoning systems, belief-boxes) allows for a different interpretation of the *modus ponens* because:

‘people do seem to be intrinsically motivated to entertain or to avoid certain types of thought or sequence of thought. Thus if someone finds himself thinking that P while thinking that $P \supset Q$, then he will feel *compelled*, in consequence, to think that Q . And if someone finds herself thinking that P while also thinking that $\sim P$, then she will feel herself *obligated* to eliminate one or other of those two thoughts. [...] The explanation is that the system for normative reasoning and belief has acquired a rule requiring sequences of thought/action that take the form of *modus ponens*, as well as a rule requiring the avoidance of contradiction, and is generating intrinsic motivations accordingly’ (2006:261)

Carruthers claims that his account can explain how people’s conscious thinking may seem to be normative.

Against Carruthers’ mechanical view on reasoning, Brandom might, in his turn, argue that it cannot explain how people acquire new concepts. For Brandom, the process of acquiring concepts, in its most basic form, follows the feedback-governed structure of a Test-Operate-Test-Exit cycle in which the knower alternately acts and perceives the results of her action, at each moment checking whether she is on her way of realising her goal.¹⁴¹ This is the case, Brandom claims, for actions as simple as reaching for a doorknob: the knower reaches for the doorknob, somehow fails (goal not reached), performs a new attempt, succeeds (goal reached and exits).¹⁴² Important to this updating

141 From a historical perspective, Brandom sees this view of learning as a TOTE-cycle as the answer that the classical American pragmatists, following Hegel (2004), came up with in response to what Brandom calls a two-phase account of acquiring concepts and which Brandom associates with Kant as much as with Carnap (2002, chapter 7). Brandom rejects this account. It assumes that concepts are already fully determined, providing as it were the ‘prototype’ (the ‘universal’, the ‘sense’) of something which only needs to be recognised once the concept is applied. In his account of the acquisition of knowledge, having concepts is already using them (i.e., there is no ‘having’ before ‘using’ them). For a congenial criticism: see Wilson (1982). Carruthers’ evolutionary nativism might be understood as also implying this two-phase approach: the brain has certain conceptual modules [algorithm] that need other mechanisms (such as observation) for their application. The conceptual modules seem to work as ‘prototypes’. See Carruthers’ discussion in (2000:55, note 18). However, as a philosopher interested in cognitive science, Carruthers here typically focuses on the recognitional function of these concepts. Such one-sided interest is criticised by Wilson, who believes that it still leaves ‘traditional philosophical problems about concepts’ unaddressed (2006:617). Carruthers denies, however, that the idea of conceptual modules commits him to what he calls ‘classical (‘definitional’) or neoclassical theory of concepts’ (2006:293, note 9). The semantic primitives he appeals to instead, nevertheless, seem to function as ‘prototypes’.

142 The doorknob example is probably borrowed from Fodor (1998). Wilson 2006:100) is critical about using the doorknob example to explore the problem of what concepts are about.

process is, perhaps surprisingly, that people make mistakes.¹⁴³ This is not proof for their lack of rationality, as some social psychologists seem to think, but instead forms an essential part of their existence as knowers ('we are discursively born into a state of sin'). Especially conceptual mistakes (in contrast to, for example, merely observational mistakes) provide the starting point of the process of acquiring *new concepts*. However, learning (the acquisition of concepts) based on *conceptual* mistakes can only be made sense of if the inferential capacity is understood in *normative* terms. Below I will focus on conceptual mistakes as what requires a normative rather than a brute-causal explanation of acquiring knowledge.

Before I move on to further elucidate Brandom's view on making conceptual mistakes, I linger for a moment and consider what might be an example of a conceptual mistake. As is always the case, choosing an example is a delicate matter. As part of his Sellarsian criticism of reliabilism, Brandom himself provides the case of someone called John who can distinguish between green and blue. After a series of experiences, John learns, however, a more specific ability: to distinguish things that, under certain circumstances (when he looks at them in electric lights), *appear* to be green (but are in fact blue), from things that are *really* green (when compared in the same daylight circumstances).¹⁴⁴ The problem for Brandom, I think, is that this example, although it exemplifies the TOTE-cycle, may still be construed as a causal, reliabilist process in Carruthers' sense. The reason for this is that it might still be interpreted as exemplifying an epistemological updating. John might still think that his concepts (even his concept of 'blue') have not changed. I think Brandom's point about the process of acquiring new concepts as a *normative* process will become clear once we move from the epistemological level to the more basic semantic level. To illustrate how the updating process may lead to the adaption (or: the development) of our concepts, I will borrow an example from Michael Liston:

'The first English-speaking colonists arrived in the New World with a stock of classificatory predicates that had unproblematic applications in the Old World. Their common linguistic training disposed them to apply many of these predicates to New World objects. Old World usage applied 'rabbit' to members of *Oryctolagus cuniculus* [sic], altricial creatures that live in burrows; New World usage applied it to members of *Lepus*, precocial creatures that live in forms. The colonists (circa 1620) are disposed to

143 Brandom (1998) criticises reliabilism by claiming that the acquisition of knowledge cannot be explained by observation alone but that it requires a second ability: reason-giving (he understands rationality specifically in a social context). He chooses to contrast the two capacities as non-inferential (observation) versus inferential (reason-giving). I think this is somewhat unfortunate as it tends to obfuscate his own point that rationality is a matter of doing. The TOTE-cycle, as I understand it, makes it clearer that the basic process of acquiring knowledge is a matter of observing and doing (and perhaps it would be better even to understand 'observing' as a kind of doing). It would mean, I think, that rationality if it is just reason-giving (just a game of reasons), could not explain making errors.

144 I borrow this example from Brandom (2002:355).

apply ‘rabbit’ to the hares they encounter in their new environs. They confidently made a natural and understandable error: based on their local Old World successes, they over-generalized the application conditions for some of their classificatory predicates’ (Liston 2010:811).

In this example, the colonists use their concept of ‘rabbit’ incorrectly but, at this stage of telling the story, they do not know about this.

Is it possible to explain, at a conceptual level, what is going on when people make mistakes, as do the colonists in the above example? Using concepts requires a variety of cognitive skills. It includes, for example, surprisingly perhaps, the capacity to ‘*ignore* some factors one is capable of attending to’ (Brandom 2008a:81). The numerous objects that knowers deal with both differ from each other and resemble each other in many respects. People need to be able, as knowers, to privilege some of these differences and similarities and ignore others, i.e. consider them as irrelevant.¹⁴⁵ It also requires that people take concepts to have a certain range of counterfactual robustness (section 6): they understand that some beast may be a lioness if it is a mammal, but that it cannot be one if it is laying eggs. It does *not*, however, require a grasp of *all* the conditions under which an inference would (not) go through.¹⁴⁶ Now people may either ignore the wrong differences (or similarities) or be wrong about certain counterfactuals.¹⁴⁷ Once they do so, they start to make mistakes. This may go on for a long time without them noticing it.

The significance of alethic modality, as we have seen (section 6), lies in the possibility it provides for expressing that one and the same object cannot have two incompatible properties. It is not possible, for example, that a mammal has the characteristic of digging burrows AND does not have the characteristic of digging burrows. If digging burrows is characteristic of rabbits, but not of hares, then the mammals that the colonists observe in the New World are either rabbits or hares. According to naturalists, even knowers can be understood as being subjected to mechanical laws. So how would it work if we described knowers, the colonists in this case, as if they were ‘objects’ having two incompatible properties, namely as having both the claim that all rabbits are burrowers AND the claim that all rabbits are not burrowers? But treating ‘incompatible claims’

145 In the short story ‘Funes el memorioso’, Borges is telling about a man who is capable, due to his extremely exceptional memory, to remember all the details he observes and who, for that reason, values them all as equally relevant. Nevertheless, as the narrator concludes, ‘no era muy capaz de pensar’ because thinking is ‘olvidar diferencias’ (2006[1944]:524).

146 This is what Brandom calls the ‘non-monotonicity of material inference’. More about this: Brandom (2000, chapter 2).

147 Mark Wilson shares some of the concerns Brandom has. He, too, opposes what he calls the ‘classical account of extension’ (1982:555) and what Brandom calls the ‘complete or maximal determinateness of concepts’ (2002:213). Wilson (1982), while he analyses what conceptual change means, claims that concepts, as they are used by a community by people, are bound to a certain range of application. This notion, as Wilson explains it, suggests that many of our concepts might be local: what Brandom (2011) refers to as a ‘microstructure’. As I understand Wilson, people may make a mistake by trespassing the boundaries of the range of application that holds for their conventional concepts.

as ‘incompatible properties’, i.e. describing them in the vocabulary of alethic modality does not work. People are often committed to incompatible claims. Apparently, having ‘incompatible claims’ is not the same as having ‘incompatible properties’. It suggests that incompatible claims are of a different nature than incompatible properties. In this respect, the vocabulary of alethic modality does not help to categorise knowers as objects.

But then what? Brandom introduces normative modality to make sense of this. To describe the process of doxastic updating, Brandom shifts from alethic modality to a different kind of modality: *deontic* modality, which helps to express the normativity of this updating process. The vocabulary of deontic modality allows us to say that it is impossible to endorse incompatible claims, but here the impossibility is not a matter of facts, but a matter of normativity: people ought not to endorse incompatible claims. It says that people should make up their minds once they are committed to incompatible claims. It does *not* say that they cannot have, as a matter of fact, incompatible beliefs.

‘If p and q are incompatible in the *alethic modal* sense, then it is necessary that not (p and q). But if p and q are incompatible in the *normative deontic* sense, then it is indeed required that one not be *committed* to (p and q), in the sense that one *ought* not to be, but it does not all follow that one *cannot* be, or is *in fact* not so committed. The sort of looseness of fit between what is necessary or required in the deontic normative sense and what is possible or actual is not even intelligible in the alethic modal sense of ‘necessity’ (Brandom 2008a:192).

Having incompatible beliefs ‘forces’ people to make a decision between both claims. But the ‘force’ of this type of incompatibility is different from causality. It is a normative ‘force’. What discursive beings do can be described if we use the vocabulary of normative modality in a sense that is not optional for objects.

As we assumed until now, the colonists made a mistake but did not know about it. Not just we, as the modern, post-colonist Interpreters, but the colonists too may, at some point, have viewed themselves as being committed to two incompatible concepts:

‘They can remain unaware of the shifting usage and of the semantic tensions in their language so long as they merely observe the passing leporidae, predict the next rabbit sighting, or hunt on the basis of superficial classification. But should they attempt other tasks (like hunting by flushing rabbits out of their burrows), they may find themselves in trouble (with a bootless hunt). As they try to diagnose what went wrong, they realize that the local rabbits do not burrow, which contradicts their belief that all rabbits are burrowers. There may follow a period of semantic agnosticism or confusion, then a corrected picture that captures the working semantics of their language’ (Liston 2010:814-815).

The discovery of their mistake, and therefore of the deontic incompatibility of some of their claims, allows them to enter the process of correcting, i.e. updating their concepts.¹⁴⁸ It would result in solving somehow the deontic incompatibility, for example, by deciding that what they previously thought were rabbits are not rabbits at all.¹⁴⁹ The deontic incompatibility between two claims can only be made sense of in terms of a normative sense of inferential rationality. And this forces the revision of one's concepts. The point is not that this person will actually do this, but just that, if the person does it, it is forced by the *deontic* incompatibility between both claims.

Carruthers would not deny, I assume, that people can be committed to two incompatible claims. He might still claim that this, and the updating process that follows from it, can be described as a causal-mechanical process. How would a modal concept like MUST, in his account, force a decision between two incompatible commitments? If he maintains his claim that the updating process is basically mechanical in nature, it will still rely on the use of the vocabulary of deontic modality. To start with, if he wants to explain in purely causal-mechanical terms how a person may be committed to two incompatible claims and how this incompatibility is resolved, he must adopt towards that person the perspective of an Interpreter in Dennett's sense. But to express what he wants to explain – i.e. the incompatibility of two claims some person endorses – he, as an Interpreter, must use the vocabulary of deontic modality. If, however, as he claims, an incompatibility ('contradiction') between two claims is nothing more than what prompts a person, in this case, the Interpreter, to 'feel herself *obligated* to eliminate' one of those claims, then how will the Interpreter know that *her* feeling of obligation is reliable? How will the Interpreter know for sure that she is not manipulated by some evil neuroscientist? As with the determinists in the previous section, he may be able to explain the world in causal-mechanical terms, but he is not able to explain that this world is inhabited by knowers. And if he cannot explain this, he cannot explain his own position.¹⁵⁰

Carruthers and Brandom both agree that people are knowers. And this means, amongst

148 This might suggest a new way of analysing transcendental freedom. Early modern theorists sometimes analysed what I call transcendental freedom as the 'freedom to do otherwise'. Notions of positive freedom may be understood as introducing yet a new of analysing transcendental freedom, namely as the 'freedom to choose otherwise'. The ability to trace incompatibilities opens the possibility of adjusting one's claims as a way of solving incompatibilities between various claims one has committed oneself. Transcendental freedom might, therefore, on this picture, be analysed as the 'freedom to claim otherwise'.

149 This picture of the updating process is not wrong but is admittedly simplistic. Wilson (1982) would argue, I think, that this is not the only possibility of predicting how the colonists will deal, cognitively, with their discovery that the animals they call 'rabbits' do not dig burrows. They might still, for example, accept that these 'hares' are simply rabbits that do not dig burrows and go on using the word 'rabbit' for hares. In this case, they need not feel that they changed the meaning of their concept 'rabbit'.

150 A different criticism is offered by Wilson. Carruthers is a 'theory-theorist' (1996). Wilson 2006:128-129) believes, however, that the 'theory' theory of concepts is 'counterproductive and obscurantist'.

others, that people may sometimes have correct knowledge and sometimes have incorrect knowledge. But they approach the issue of having (in)correct knowledge from opposite perspectives. Carruthers basically provides an ontological, more specifically neurological, account of having correct/incorrect knowledge. The underlying strategy is one of explaining concepts in terms of their *causal* roles.¹⁵¹ Brandom instead attempts to understand them in their *normative* roles. The difference is that Brandom's 'person' has a normative attitude, whereas Carruthers' 'person' is a being whose reasonings obey the Natural Unconscious. Nevertheless, they are both able to answer the question. But Carruthers' answer, as should be clear by now, leads to a useless explanation of 'making mistakes': to describe ourselves *not* as fallible, normative (rather than fallible, merely dispositional, i.e. non-discursive) beings is really committing cognitive suicide. And so, the Manipulation Argument can be explained in a different sense: *transcendentally free, but not responsible*.

5.8. Conclusion

In this chapter, which functions as the turning point in my overall argument, I have argued against the determinism thesis. To criticise it, I introduced two ideas: that freedom is a two-level notion and that rationality is inferential. The first idea helped to open space for an alternative, not yet considered option in the standard debate on determinism: performing free actions but not being responsible for them. The second idea was derived from Robert Brandom's normative-pragmatic view on what it means that people are *knowers* or, in more traditional philosophical terms, *rational beings*. In his view, rationality is, at the most basic level, practical in a non-deliberative way ('meaning is use'). Only at a more sophisticated, linguistic level does deliberation arise, which can help to express what is rational at a basic level. Even at this basic, practical level, rationality is inferential in the sense that people's actions display their grasp of certain concepts, of what can be inferred from what, what is a reason for what (the space of reasons). From a young age onwards, people learn to master many concepts by trying, trying again and making mistakes (the Test-Operate-Test-Exit cycle). This implies that modality is a feature, not just of people's reasoning but also of their practices and of the world itself. Finally, part of Brandom's view on inferential rationality is that we should understand grasping concepts as a normative affair: having concepts is authorising oneself to what follows from one concept for other concepts. The idea of inferential rationality was aligned with the idea of freedom as a two-level notion. I have shown how both this alignment and the idea of people as knowers (in a normative-modal-practical sense) undermine the determinism thesis.

My aim was to show that the mentalist framework, to the extent that it implies the determinism thesis (which I think it does), is deeply problematic because it is not able to explain what it means that people are *knowers*. At the same time, I wanted to develop an alternative, Brandomian-pragmatist framework. In the next chapter, I will further elaborate on this framework by suggesting how it can be used to conceptualise the veil of prejudice.

¹⁵¹ Wilson, too, is critical about views that understand concepts in terms of their causal role (see 2006:611). I think this criticism also holds against Carruthers.

6

Chapter 6

A pragmatist notion
of the veil of prejudice

6.1. Introduction

My aim in this chapter is to explore how Brandom's notion of inferential rationality might be used to make sense of the veil of prejudice. This notion involves two dimensions that may pose difficulties to developing a Brandomian-pragmatist account of hidden prejudices: its possibly *perspectival* (or: *scorekeeping*) and its *voluntarist* dimension. The second dimension accounts for the difficulty of theorising the *hidden* nature, the first for the difficulty of theorising the *social* nature of hidden prejudices. I will therefore focus on these two questions: first, what does it mean that the prejudices people have, are *hidden* from them? And second, what does it mean that these people *share* their (hidden) prejudice?

In the next section, I will begin by explaining how, in Brandom's view, inferential rationality is rooted in *social* practices (section 2). To analyse them, Brandom uses the basic notion 'attribution' (or: 'scorekeeping'): people attribute to other people and to themselves commitments (Brandom's technical terms for beliefs and intentions). His notion of attribution can also be used at a meta-level: we might imagine how, for example, a sociologist (or, for that matter, an anthropologist) observes a community and tries to make sense of its practices by attributing hidden prejudices. For the sections that will follow, I will therefore use Brandom's view as a kind of sociological theory. In my analyses, I will sometimes introduce the figure of an outsider that attributes hidden prejudice to people within a society. I will call her the sociological Interpreter. In an intermezzo, I will briefly attend to the sociological problem of understanding agency/structure as an equivalent of the veil of prejudice. In this context, I will explore some of the considerations that led Bourdieu to develop his notion of habitus (section 3). While Brandom has focused largely on what is made explicit in linguistic practices, rather than what is implicit in practices of rationality, Bourdieu instead has largely focused on what is implicit in social practices. As he too rejected the mentalist framework, he had to deal with a similar challenge of theorising the rationality of social practices. His notion of habitus can also be interpreted as fitting into a pragmatist solution. It is interesting, therefore, to see how Bourdieu, as a sociologist, discusses the various options for sociologists to theorise the rationality of social practices, what problems he sees in these various options and how the notion of habitus is supposed to solve or evade these problems. Bourdieu's insights will then be used to interpret Brandom's notion of rationality as a social practice in a way that retains the notion of 'authorisation' without dissolving it in a voluntarist-like notion of 'choice'.¹ For this, I will choose Brandom's notion of 'implicit' and make a proposal to understand hidden prejudices

1 Bourdieu (1972/ 1977) has been critical of the phenomenological tradition, while Brandom (2004) has critically reflected on pragmatism. At the same time, their views can be situated in both traditions. Combining their views would have particular philosophical relevance at the methodological level because it offers the opportunity of reconciling the phenomenological and pragmatist tradition. But this I will not explore here.

as implicit commitments (section 4). I will also use arguments that Brandom himself has already developed and which may prove useful to a sociological interpretation of practices of prejudices as hidden. One issue remains: how to understand that people share their prejudices, hidden to themselves, with each other? Analytical philosophers have addressed a similar problem as one of either *collective action* or *common knowledge*. I will briefly discuss how they have analysed this problem and then show how Brandom's account of scorekeeping presents a different and, I think, better solution.

In chapter 2, it was argued that practices of (hidden) prejudices consist not only in practices of discrimination but also in practices of silencing. Integrating this insight in the Brandomian-pragmatist version of the veil of prejudice will be deferred to the next chapter (section 4) as this requires, in my view, a notion of responsible agency.

6.2. The space of reasons as a social space

Brandom's account of rationality, as was already suggested in the previous chapter, is a *social* account: while semantics is understood in terms of practices ('meaning is use'), it should even more so be understood in terms of *social* practices. People become rational beings, not just based on their biological potentiality (e.g. having a large brain) or their capabilities to learn from their senses (according to a Test-Operate-Test-Exit cycle), but because they are raised in a community of rational beings. To be rational is to be rational in the company of others. In this section, I will provide an outline of Brandom's account of rationality as a social practice, one that will be terse and simplified, but which will suffice to explore, in the next sections, whether his account can be used to incorporate some notion of hidden prejudices. Below I will first briefly explain in what senses Brandom's account is different from other social accounts of responsibility. Then I will move on to explain his account in more detail. For this, I will introduce some of the terms that Brandom's uses for analysing practices that are rational in a broad sense, linguistic and non-linguistic. An important part of Brandom's work is devoted to showing how *linguistic*, social practices, in their turn, make it possible to understand linguistic meanings and more broadly intentional contents. For this, he employs a more specific terminology ('assertion', 'speech act', 'ascription', et cetera). I will not get into this as this is not needed to clarify some of Brandom's basic ideas about rationality as a social practice or for the argument in the next sections.

Social accounts of responsibility share the basic idea that responsibility is a matter of understanding 'responsibility' in terms of social practices (of 'holding responsible') rather than in terms of the right kind of causal agency. They are reminiscent, therefore, of the early modern tradition of imputation. People are responsible, not so much because they choose to be responsible, but because they are held responsible by others: the assessment of responsibility relies on the attribution of responsibility by others. Within analytical,

Anglo-American philosophy, accounts of *responsibility as a social practice* seem rare.² Two important and influential proposals, however, have been put forward by Strawson³ in his 'Freedom and Resentment' (1962) and Dennett in various of his writings.⁴ What distinguishes their accounts from Brandom's, is that these are still embedded within a naturalistic perspective.⁴ Below I will explain how Brandom understands the space of reasons as a *social space*.

Brandom observes that practices of normative bindingness can be thought of as *social* in at least two distinct ways: as either an I-thou or an I-we sociality.⁵ Brandom claims sociality has basically an I-thou sociality structure to which the exchange of reasons is essential or, as Brandom calls it: the game of giving and asking for reasons. Yet another way of putting it: Brandom's notion of rationality is perspectival.⁶ It implies that what is rational is not simply the result of one monological perspective but of different, dialogical perspectives which connect with each other as first-person perspective versus second-person perspective. It suggests that important to his account is the second-person perspective.

It is quite standard for philosophers (at least within Anglo-American philosophy) to analyse knowledge in terms of *justified true belief*. A person can be said to know *x*, if he believes *x*, has a reason to believe *x* and, finally, *x* is true.⁷ For various reasons, Brandom favours a different terminology. First of all, he replaces 'belief' with his term 'commitment'.⁸ He uses this term in a broader sense because a commitment may either be doxastic or practical: while the first corresponds to 'belief', the second is the commitment to act.⁹ If a person has a reason to have a commitment, differently put: if this commitment functions as the conclusion of an inference, she has an entitlement to that commitment. This corresponds with 'justified'. In the previous chapter, I have already indicated that Brandom understands meaning in terms of use. While many philosophers who analyse knowledge are interested in the question of what the conditions

2 See discussions of several accounts of responsibility in Smiley (1992, esp. part two). The situation in the continental tradition is probably not much different: see, e.g. Raffoul (2010). Sartre is often cited as a representative of those committed to individualistic accounts of responsibility. Levinas can perhaps be viewed as his counterpart (cf. Raffoul 2010). Exceptions are, e.g. Smiley (1992) and Walker (2007).

3 Strawson's account has influenced, amongst others, Fischer and Ravizza (1998).

4 Brandom (1994, chapter 1) explicitly discusses Dennett's approach on 'original intentionality' but rejects his naturalist turn.

5 See Brandom (1994:62) and (1994:716).

6 This perspectival view does not necessarily imply some form of moral relativism. It goes beyond my present scope to explain this, but: see Bader (2007, chapter 2) for a discussion on how moral relativism may be avoided.

7 For a classic discussion of justified true belief: Gettier (1963).

8 The trouble with 'belief' in ordinary but also philosophical English is that belief has an ambiguous meaning. It may refer to both what Brandom calls 'empirical belief', a belief a person actually has, and normative belief, a belief to which a person should be committed, given certain other beliefs she actually has (and endorses).

9 See Brandom (2000:83).

of truth are, Brandom follows a different meta-strategy: he discards this question and instead understands ‘truth’ in terms of an attitude, i.e. of adopting an attitude of taking x as true, or, also in Brandom’s terminology, of undertaking a commitment.¹⁰ So ‘belief’, ‘justified’ and ‘true’ are replaced with respectively ‘commitment’, ‘entitled’ and ‘undertaking (i.e. a commitment)’. Commitments and entitlement together are what Brandom calls ‘deontic statuses’.

Imagine a world resembling earth that is inhabited by only one rational being, perhaps the last survivor of a species of rational beings. With no one left to talk to, the last survivor is still doing things that indicate her rationality. If she sees some animal running through the field and thinks that it is a fox, it commits her also to claiming (normatively, not empirically) that it was a mammal she saw running. If rationality is explained like this in *intercontent* relations, i.e. in terms of inferential relations between conceptual contents (e.g. between ‘fox’ and ‘mammal’), it is tempting to understand normative bindingness as a process that somehow takes place within some interior space: a person’s mind, her subjectivity, as if people’s knowledge was a private issue, as if no other people are needed to explain knowledge. In Brandom’s view, having commitments and entitlements is not merely an intercontent, intrapersonal relation but also an *interpersonal* relation. ‘Interpersonal’ might, however, still be understood in various ways. In the following paragraphs, I will introduce step by step a new type of interpersonal relation, starting with a minimal I-him/her type, until we arrive at what Brandom considers to be the I-thou type. While the order of explanation as presented here proceeds from a ‘minimal’ to an I-thou type of interpersonal, it should be reminded that in Brandom’s view, the *conceptual* order is really reversed: it is only due to the I-thou structure that an I-him/her structure is possible.¹¹

For many games, *scorekeeping* is important because it allows players of the game (and their audiences) to keep track of the score at any stage in the game. What the score is – what counts as ‘scoring’ – will depend on various rules and often, as in unclear cases, on what an umpire says the right score is. Brandom uses ‘scorekeeping’ as a metaphor for explaining sociality.¹² The activities it consists of are necessary for every type of sociality that I will introduce below. The central activity is ‘attributing’. If a person undertakes a commitment, according to Brandom, this is not something that can be understood by

10 Also, see chapter 5, section 5.

11 As Brandom puts it, somewhat technically: ‘The conceptual contents employed in *monological* reasoning, in which all the premises and conclusions are potential commitments of one individual, are parasitic on and intelligible *only* in terms of the conceptual contents conferred by *dialogical* reasoning, in which the issue of what follows from what essentially involves assessments from the different social perspectives of interlocutors with different background commitments. Representationally contentful claims arise in the social context of communication and only then are available to be employed in solitary cogitation’ (1994:497).

12 Brandom borrows the metaphor from Lewis (1979), who took it from the game of baseball for an analysis of conversation ‘or other process of linguistic interaction’ (1979:344). However, Brandom’s notion of ‘scorekeeping’ is different from how it is used in actual games because ‘scorekeeping’ does not involve any ‘official’ score: each scorekeeper keeps her own score.

itself. Undertaking a commitment is ‘doing something that makes it appropriate for others to *attribute* it’ (Brandom 2000:174).¹³ If a scorekeeper attributes some commitment to another scorekeeper, it may be the case either that she had already undertaken this commitment (taken it to be true) herself, or, persuaded perhaps by what the fellow-scorekeeper was saying, she decides to take over this commitment and undertake it herself. But in all cases, the attribution of a commitment implies that scorekeepers are able in principle to distinguish between commitments they attribute and those they undertake themselves, i.e. to distinguish between another person’s perspective and one’s own perspective.¹⁴ The notion of ‘attribution’ allows us to introduce the first, minimal type of rationality: an I-him/her sociality. While all the fellow-beings of the last survivor have died, let us imagine yet another being, a scientist, who has come from a distant planet, on her own, to study this last survivor and to learn more about his civilisation. The relation between the scientist and the last survivor can be analysed as a unilateral relation between an attributor and an addressee. Because the scientist in observing the last survivor hides from him, the scientist and the last survivor have no I-thou relation in Brandom’s sense. The last survivor does not know about the scientist.

The scientist who attributes all kinds of commitments to the last survivor she observes undertakes a commitment herself as well. In attributing commitments to the last survivor – commitments she thinks that the last survivor undertakes –, at the same time, she *undertakes* commitments herself because she commits herself to the attribution of commitments to that other person. Attributing a commitment is, therefore, also doing something that makes it appropriate for others to attribute it. But in the case of my example, these attributions cannot be attributed by the last survivor as the scientist remains hidden from him. The space of reasons, understood as a place where several rational beings are present, is, however, a *public* space. Brandom’s point about I-thou sociality is that *all* rational beings are in principle scorekeepers and therefore attributors. They have to be in a position, therefore, in principle at least, of being able to mutually attribute commitments to each other. So, we will change the scenery of the last survivor and the scientist to one where people are able to observe each other. Let us imagine this time, but still science-fictional, a situation where people live all alone in different places, have no means of communicating with each other (improbable as this science-fictional world is, but I nevertheless assume it to be this way) and can observe each other due to all kinds of technical devices, such as cameras and microphones. It would resemble somewhat Bentham’s panopticon, but it would be a kind of democratised panopticon:

13 At this point, Brandom seems a similarity between his view and that of Dennett because attributing a commitment is, as Brandom says, ‘an implicit version of adopting the intentional stance’ (1994:61).

14 Failing to distinguish between attributing and undertaking a commitment is what Brandom believes lies at the heart of verificationist epistemologies: ‘The classical metaphysics of truth properties misconstrues what one is doing in endorsing the claim as *describing* it in a special way. It confuses *attributing* and *undertaking* or *acknowledging* commitments’ (1994:515). Also, see my discussion in chapter 4, section 5.

everyone observes everyone. Each can keep track of other people's commitments (and perhaps entitlements).

Animals make mistakes, and human beings make mistakes. But human beings make mistakes in a different way: as concept-users. In the previous chapter (section 5), we have seen that people, as rational beings, are concept-users and that they, as concept-users, bind themselves to concepts. The very idea of being bound by a concept, at the heart of Brandom's account, is precisely that in binding oneself to a concept, one binds oneself to what is implied by that concept, i.e. to the inferential relations of that concept with other concepts. Although a person may choose to bind herself to a certain concept, she cannot choose what follows from that. Once she is committed to, for example, the concept of 'fox', she is committed to claiming that a fox is an animal on four legs with a brushy tail. This 'binding' part of being a concept-user is what Brandom calls 'objectivity'. It allows a scorekeeper to distinguish between the *objective* content of a concept and the *subjective* application of that concept or, in Brandom's own technical vocabulary, between *deontic statuses* (i.e. commitments, entitlements) and *deontic attitudes* (undertakings or attributions of commitments or entitlements). Although deontic attitudes can be said to institute concepts – some nature lover takes it as correct for example that foxes have brushy, white-tipped tails –, this does not imply that any 'taking this-or-that as correct' is in fact correct. People may be mistaken. A scorekeeper might acknowledge that people are committed to the concept of 'fox' but are mistaken in some aspect. Within the democratic panopticon, for example, a biologist might notice how a nature lover strolling through the fields sees an animal running on four legs with a brushy tail, saying to herself: 'I thought it was a fox, but I see it has a black-tipped tail, so it must be a different animal'. The biologist may then realise that the nature lover does not know about the gray fox who has a black-tipped tail, in contrast to the red fox who has a white-tipped tail.

So far, we have seen how scorekeeping roughly works. The metaphor works to explain how a scorekeeper may observe differences between her own perspective and the perspective of another person (the attributee). The type of sociality as it occurs within the world of the democratic panopticon, where everyone can keep track of each other's commitments, is not, however, the I-thou sociality Brandom aims for. The metaphor of 'scorekeeping' is exactly missing the aspect of a communicative interaction between a first-person and a second-person perspective. All that the scorekeepers can do is observe each other and keep scores. It seems the next step would be to introduce communication between scorekeepers, for example, as conversation. But how does rationality connect with conversation? Simply assuming that the people in the democratic panopticon would be able to communicate with each other does not yet explain why conversation is needed for rationality of the sort Brandom argues for: rationality with an I-thou structure. What in Brandom's account of rationality requires the giving and asking for reasons? Why does rationality imply that people give and ask for reasons? Habermas

(2000), for his part, claims that the point about rationality, specifically understood (by both Brandom and Habermas) as *linguistic practice*, is about mutual understanding (*Verständigung*). Brandom refuses, however, to project any such orientation on people. Brandom disagrees with Habermas, not because he thinks that ‘linguistic practice has some *other* point, but because I think it is a mistake to think of it as having a point at all’ (2000:363). There is no categorical imperative to be rational. There is no purpose to linguistic rationality. Brandom acknowledges that rationality enables people to pursue ends in a rational way (instrumental rationality). But agreement (or some other end that is attributed to linguistic practice) can be made sense of only *within* the context of rational practice, *not in advance* of rational practice as such (in Brandom’s broader sense, i.e. what I called in the previous chapter ‘theoretical rationality’).

Brandom wants to steer away from teleological explanations of rationality.¹⁵ Apparently, he chooses a different strategy for linking rationality (linguistic practice) and conversation. But which one? Brandom tries to explain how practices of rationality work. He wants to figure out what a community of beings *ought* to do for their practices to count as rational. The ‘ought’, as Brandom uses it, is internal to the concept of rationality. Anyone who claims to know what a fox is (to have mastered the notion of ‘fox’) and sees a fox, ought to claim that *that* animal is a fox if she is to count as knowing what a fox is. In a similar way, any community who claims (or who is claimed by an Interpreter) to perform practices of rationality ought to perform such-and-such actions to count as a rational community. Part of this ‘ought’-story, as we have seen, is that people undertake commitments and are entitled to them. But does it mean that people ‘ought’ to communicate with each other? And if so, in what way? People may talk about anything they want to – the weather, politics, work – and they may have such talks for all kinds of reasons – enjoying company, curiosity, have a quarrel, settle a fight. From the perspective of a theory of rationality, what matters about analysing ‘conversation’ is how it sustains the practices of inferential rationality. From this perspective, what counts is that conversation ‘has the significance of communication only where the commitments of speakers and audience differ’ (1994:511). For Brandom, this means, if I interpret him correctly, that conversation has the important function for rationality to uncover and possibly correct mistakes. Conversation is what enables the biologist and the nature lover to discuss a possible mistake, one that is, in this case, attributed by the biologist to the nature lover (mistaking a certain animal with a black-tipped tail for not being a fox).¹⁶ It is not needed, however, that people engage

15 Habermas’s views belong to the teleological approach. Brandom also rejects the evolutionary explanation of rationality. Gauthier (1986) provides an example of a theory of rationality that may be counted perhaps to both approaches: people (as a species) as having evolved to ‘constrained maximizers’.

16 It may, of course, be the case that the nature lover knows very well that the gray fox is a fox, but that he has learned, due to whatever idiosyncratic study book, to use the term ‘fox’ only for red foxes, whereas he uses a different term for the notion ‘fox’. But a conversation may, of course, bring this to light, or even other scenario’s that may explain the differences in perspectives between the biologist and nature lover.

in conversation because they intend to correct a mistake. Conversation may function to uncover mistakes even if people have very different intentions with their conversation.

A few warnings are in place. First, it should again be reminded that Brandom's notion of rationality is not an intellectual one. When people communicate, they are constantly involved in some form of scorekeeping: whatever the other person said (is believed to be saying) is compared to one's own commitments. But this scorekeeping is, as Brandom explains, 'implicit, practical interpretation, not explicit theoretical hypothesis formation' (1994:508).¹⁷ Furthermore: Brandom does *not* claim that all actual conversations will show differences in perspectives between people or, if they do, that they will always result in correcting different perspectives or, if they do, result in making the right corrections of mistakes.¹⁸ What Brandom will claim, I think, is that *without conversation*, or perhaps better, *without communication*, rationality cannot ever come off the ground.

If conversation functions to uncover differences between scorekeepers, it may still come in different varieties. What variety does Brandom have in mind? One claim might be that the right kind of I-thou conversation requires questioning *every* commitment that a scorekeeper attributes to another person. But this 'threatens a regress on claim contents. At each stage, vindications of one commitment may involve an appeal to commitments that have not previously been invoked, for which the issue of demonstrating entitlement can arise anew' (1994:176). Radical challenging of each other's commitments would make it impossible, however, for practices of inferential rationality itself to come off the ground. And as a result, not even 'radical challenging' itself would be possible. To analyse why social practices of inferential rationality do nevertheless exist, Brandom takes them to have a *default and challenge* structure.¹⁹ When they attribute commitments to another person, people *also* attribute to her entitlements to that commitment (i.e. attribute to her the ability to justify her commitment). This is what Brandom calls attributions of *entitlement by default*: we usually do not ask people to vindicate their commitments. This does not mean, however, that commitments are

17 One of the big challenges for Brandom's project, as Loeffler (2018) sees it, is to explain that norms are implicit in what people do (see my discussion in chapter 2, section 6), but also that attributions themselves may be implicit. Loeffler thinks it important for Brandom and other neo-pragmatists to provide such an argument, especially in response to adherents of the 'theory of mind' (see Carruthers and Smith 1996) who have made proposals to make sense of the same phenomenon (people understanding other people), but in terms of non-normative, naturalistic mechanisms. In section 5, I will take up this challenge to some extent.

18 While an individual may be mistaken, the community as such may also be mistaken, even about concepts it has committed itself to. I would claim that, as a minimum, a community cannot be said to be mistaken about everything (analogous to what was discussed in chapter 4, section 5). The intriguing question would be: how many mistakes can a community make or, putting it the other way, how many conversations of the right type should a community have to remain a community that performs practices of rationality? When is it still rational, and when does it lose its rationality? My guess is that Brandom would claim that this is an empirical matter and that philosophers cannot tell in advance where exactly the critical boundary of making mistakes is.

19 Also, see Brandom (1994:176-178).

immune to criticism. A radical immunity of commitments (putting into doubt the entitlement to *any* commitment) would make it impossible for practices of rationality to get off the ground. Commitments may be *challenged*, however, if the challenger has a reason to question another person's commitment, i.e. if she is *entitled* to the challenge. In that case, conversation may function for all the participants in the conversation to trace the differences in commitments and entitlements and to explore what is the right commitment.

Whether or not, during a conversation or any other form of communication, an entitlement arises for a scorekeeper to *challenge* her antagonist's commitments depends on how the communicative process proceeds. It might be claimed, for example, that what is still only in the speaker's mind *before* communicating is something speaker and listener share *after* successful communication. In this simple model of communication, the speaker is presented in an active role while the listener has only a passive role, merely receiving, so to speak, whatever message the speaker gives to her. But Brandom understands listeners as having to do much more than just that. Being a listener (Brandom himself usually talks about 'audience') is not a passive role. The listener is a scorekeeper. As Brandom understands commitments in terms of their inferential roles – their role as conclusion or premise in inferences –, he analyses scorekeepers in being engaged, implicitly, in establishing the *inferential significance* of the commitments they attribute to their fellow-scorekeepers. So, a scorekeeper will constantly try to keep track, downstream, of the *inferential consequences* of an attributed commitment – what would follow from this commitment? – and, upstream, of the *inferential antecedents* – what prior commitments are needed for being entitled to the attributed commitment? This cannot be done by focussing on just the attributed commitment, as inferential consequences usually depend on a set of commitments rather than one commitment. The scorekeeper will therefore have to invoke certain other commitments that could and should function as collateral premises. These are the commitments that belong to what Brandom calls her 'background beliefs': the commitments a scorekeeper herself undertakes. These background beliefs help her to make sense of what the speaker is communicating. Although the scorekeeper, by default, accepts the commitments (and entitlements) she attributes to the speaker, she may realise at some point during the conversation that the speaker makes a mistake: he makes a claim that is not correct. Brandom uses the example of Fred and Wilma, who have a conversation at some party (1994:489–490).

Fred says: 'The man in the corner with champagne in his glass is very angry.'

What Fred does not know because he has no clear view of the corner, nor of the contents of the glass, is that the man in the corner, Barney, has ginger ale in his glass and that standing in that corner is yet another man, Nelson, who has champagne in his glass.

Wilma has better eyes and a better view of the corner. She attributes to Fred the following commitment:

Fred believes that Barney is very angry.

Wilma, therefore, uses some of her background beliefs as collateral premises to transform what Fred is saying into a meaningful commitment she may attribute to him. More complicated examples could be added. But what I hope this simple example at least shows is how scorekeeping functions in communication (as practices of rationality) that Brandom understands as I-thou sociality.

Even if it is beyond the scope of my present discussion to provide a full, detailed argument, I want to conclude this section by suggesting that scorekeeping should be understood, at least partially, as itself bound to certain social rules, i.e. as fulfilling a social role. In one tradition of personhood, people are said to perform (social, legal) roles in society. This has relevance to understanding their freedom at the contrastive level (see chapter 5, section 2). Do the roles they fulfil agree with the freedom they are entitled to? Are they somehow constrained in fulfilling their roles? Are they coerced to fulfil other (social) roles? Raising such questions are all relevant for discussing freedom at the contrastive level from a particularly political perspective. Hobbes was familiar with this notion of personhood and applied it in the contrastive sense.²⁰ Brandom's account of transcendental freedom implies, as I understand it, that performing a (social) role is also constitutive of being a person at the disjunctive level: we can only be a person (a rational, a sapient being) if we perform our conceptual activity from a social perspective, i.e. in performing *social* roles (to put it differently: rationality is necessarily perspectival). From Brandom's own writings, it may not be immediately clear that scorekeeping is doing something from a *social* perspective, as opposed to simply one's own *individual* perspective. Scorekeepers are clearly *individuals* and not, for example, communities or social groups. It is also clear that the commitments and entitlements they keep track of are attributed to other *individuals* and not to communities or social groups. But does it follow from this that individuals keep scores *as* individuals? Suppose, for example, that two children are playing soccer and one child kicks the ball into the bushes. While the first child shouts that it must have fallen into a nearby ditch, the second child sees the ball lying in the bushes and concludes that the first child is wrong. This case of an observational report may perhaps be a genuine example of an individual undertaking a commitment *as* an individual. But my claim would be that most cases can be analysed as situations in which we undertake commitments as part of a role we play, as part of a specific social perspective we take. To illustrate how undertaking a commitment binds us to what follows from this, Brandom often uses examples that are neutral with respect

20 See, e.g. Thiel (2011).

to who is saying it, as in the example of seeing a fox. But Brandom is nevertheless sensitive to the social dimensions of scorekeeping:

‘Logicians typically think of inference as involving only relations among different propositional contents, not as also potentially involving relations among different interlocutors. However, discursive practice, the giving and asking for reasons, from which inferential relations are abstracted, involves both *intercontent* and *interpersonal* dimensions. [...] [T]he issue of what follows from what essentially involves assessments from the different social perspectives of interlocutors with different background commitments’ (1994:496–497, original emphasis).

Social perspectives are manifold: parent, employer, employee, citizen. And probably the least understood, or the least recognised *social* perspective is the individual perspective. Once scorekeeping is understood as an activity that essentially takes place from certain social perspectives, this also raises questions for Brandom’s account of scorekeeping that have already been explored within the social sciences. Some of them are relevant to my discussion, such as understanding the perspective of scorekeepers as both social and unconscious. To understand the problems social scientists have with this, we now turn to Bourdieu.

6.3. Bourdieu’s ‘learned ignorance’

Brandom’s normative pragmatics provides, as we have seen in the previous section, a strategy for explaining human practices in a normative sense: people are concept-users who are engaged in social practices of grasping each other’s commitments by scorekeeping. But how does the veil of prejudice fit in? Hidden prejudices illustrate the phenomenon of dissonance: somehow, people do not have, barely have, or perhaps resist having, cognitive access to the dissonance between what they say and what they do. Instead, they interpret their own practices not for what they really are (e.g. as prejudiced), but in a different way. Does Brandom’s theoretical framework allow for theorising this dissonance? In this section, I will briefly describe how sociologists treat this dissonance as the structure/agency problem. Next, I turn to Bourdieu, as a possible ally of Brandom, and his criticisms of various options that sociologists seem to have of solving the structure/agency problem. Bourdieu, too, as does Brandom, does not rely for his own theoretical framework on the conscious/unconscious vocabulary and instead adopts what might be called a pragmatic-hermeneutic approach, relying on his notions of habitus and amnesia. In the next section, I will argue that the notion of norms implicit practices lends itself also to a pragmatic-hermeneutic approach.

Practices of prejudice are not simply the aggregated effect of various individual

people who happen to have the same prejudice but are somehow more or less systematic effects of the fact that these prejudiced people belong to the same social groups and reproduce their prejudices, even more so, perhaps, that the reproduction of these prejudices is somehow related to the distribution of power within a society to which people belong who have prejudices and those who have those prejudices. The reproduction of both prejudices and relations of power to which they contribute may be understood as following certain patterns. But do people themselves have knowledge of these patterns? This presents a familiar problem to sociologists. They have sometimes framed this problem as one of structure/agency (or: macro/micro, society/individual, system/life-world, script/role).²¹ They have generally acknowledged that the participants in a community appear not to be conscious of this structure or, if they are conscious of certain norms that govern their social practices, at least they are not conscious of deeper structures that regulate their society, structures that are, to put it differently, hidden to them. This amounts to a paradox: on the one hand, the participants in a society perform activities that, on a macro-level, display certain structures; on the other hand, they have no conscious knowledge of these structures or only distorted knowledge. This seems to be the sociological equivalent of what is the concern of our analysis: the veil of prejudice. Sociologists have sometimes taken recourse to the vocabulary of (un)consciousness, but have also at times been dissatisfied with it, criticised it and even dismissed it, usually for reasons that are closely tied to the problem of understanding the structure/agency distinction in sociological theory.²² Brandom's account of social practices as scorekeeping can be interpreted as displaying what resembles the sociological structure/agency-distinction. What Brandom analyses as scorekeeping may be understood, from the perspective of each individual, as people's agency. What Brandom understands as the norms implicit in the practices of a community, from the perspective of an outsider, a sociological Interpreter, may be understood as the equivalent of what social scientists sometimes call the 'structure' of society.²³ Brandom does not address the distinction between 'norms in practice' and 'scorekeeping' in a way that sociologists address the distinction between structure and agency. Bourdieu does.

We have already seen (in chapter 3, section 5) that Bourdieu has developed the notion of habitus and used it for a critical analysis of practices of censorship. Developing

21 On the structure/agency problem and on the problems with framing the problem in terms of such dichotomies, much literature has been written. But see Knorr-Cetina (1981).

22 For a rare and interesting, general discussion of the significance of the vocabulary of 'consciousness' for the social sciences: see Hodgkiss (2001).

23 Brandom does not refer to sociologists. He does refer to Hegel as a philosopher who has tried to capture the objective structures of societies or what Brandom has called, in Hegelian fashion, the 'social synthesis of objective spirit' (1979:187). Also, see Brandom (2004:252): 'In my view, Hegel's social recognitive story about the nature and origins of conceptual normativity was the first thoroughly non-subjectivist theory of the subject, the first to understand consciousness as a social achievement taking place as much outside what is immediately present in the particular mind as it is to be found there'.

this notion and even choosing the term ‘habitus’ was already part of Bourdieu’s development towards a pragmatic approach. As notion and as term ‘habit’ occurs in several traditions within the history of philosophy and sociology.²⁴ It recognises that people’s motivations for acting are somehow hidden, even from themselves. It accounts for the hidden nature of individual behaviour. On the other hand, it links individual behaviour to social practice. It acknowledges that the habitual practices of an individual have meanings that are beyond an individual’s purposes and can be understood by the people of a social group. While habit was once the central notion of a venerable tradition of ethical thought, originating in Aristotle, it became incorporated, ever since the nineteenth century, in a naturalist vocabulary and was taken to be some sort of psychological mechanism.²⁵ Later, various theorists have explored the links between habit and the psychoanalytic unconscious.²⁶ Bourdieu, however, developed the notion of habitus as an alternative to the notion of ‘unconsciousness’.²⁷ He had his reasons for staying away from a naturalistic conception of ‘habit’.²⁸ He, therefore, choose a slightly different term, habitus, as he explains:

‘One of the reasons for the use of the term habitus is to wish to set aside the common conception of habit as a mechanical assembly or preformed programme, as Hegel does when in the *Phenomenology of Mind* he speaks of “habit as dexterity” (1972/ 1977:218, note 47).

In developing his notion of habitus, he also wanted to resist views on human action that were too intellectualistic.²⁹ So Bourdieu introduced his notion of habitus to mark a seemingly impossible theoretical position: against standard views that understand rational actions in terms of conscious, deliberate choices and against standard views that understand non-rational actions in terms of unconscious mechanisms.

Bourdieu was led to developing his notion of habitus by his early experiences as a researcher in Algeria, performing fieldwork in Kabyle society. But he also converted his experiences into a reflection on the relation between social scientists and the object of

24 For various contributions on the history of the notion of habit: see, e.g. Nathan Brett (1981); Camic (1986); Sparrow and Hutchinson (2013); Livingston (2015).

25 See, e.g. Lockwood (2013).

26 See e.g. De Lauretis (2000); Colapietro (2000); Sullivan (2006).

27 In an interview, Bourdieu stated explicitly that consciousness and unconscious are ‘the very alternatives that the notion of habitus is meant to exclude’ (1990:10).

28 For discussions of Bourdieu’s habitus in relation to other notions of habit: see Crossley (2013a) and (2013b); Maton (2012); Wacquant (2016). For a defence of the notion of habitus as compared to ‘other concepts of normatively and rationally guided patterns of conduct’: see Pickel (2005). Ngo (2017) considers Bourdieu’s notion of habitus as possibly supporting her theoretical framework for a phenomenology of racism, but she rejects it in favour of Merleau-Ponty’s notion of habit.

29 Other theorists, too, have chosen the notion of habit to oppose intellectualistic views on human agency: see Dewey (1922); Pollard (2008). They both, however, remain committed to a naturalistic outlook.

their research: the social world. It prompted Bourdieu to undertake a criticism against the conscious/unconscious distinction as social theorists use it as a conceptual tool to capture people's agency. His criticism is motivated, at a deeper level, by a concern with a proper understanding of the social world, an understanding that should take account of the 'illusion of naturalness' that is inherent in the social world: people take their social world for granted, i.e. experience it as having a 'self-evident and natural character'.³⁰ In his reflection, Bourdieu distinguishes different possible stances ('modes of knowledge') that social theorists may take towards the social world. The two main positions are subjectivism and objectivism. The objectivist position can be split up into either a mechanistic or finalist position. I will briefly discuss these four positions and then show how the conscious/unconscious distinction, as Bourdieu believes, imposes limits on these positions that prevent them from properly explaining the social world.³¹

The social world comprises the social practices, and the primary understanding people have of them, i.e. their representations of their own practices.³² A particular feature of the social world is that people experience their world as having a 'self-evident and natural character' (1977:3), an insight Bourdieu shares with the phenomenologists. One stance towards the social world and particularly this experience, is described by Bourdieu as subjectivism. It does not yet present a social-scientific stance towards the social world. Bourdieu associates it predominantly with the position of phenomenologists (especially Sartre) but also of ethnomethodologists (e.g. Garfinkel). In the phenomenological approach, the social world is analysed by 'bracketing' the everyday experiences (the phenomenological reduction, the *epochè*), i.e. the 'natural standpoint' of people in the social world. It is criticised by Bourdieu because it sides too much with the perspective of agents and underexposes the significance of social structures. Phenomenologists exclude from their analysis the question of what the social conditions for people's primary understanding of their social world are as self-evident and natural.³³ Neglecting the question of the social conditions, phenomenologists must take for granted themselves what appears to people in their social world and cannot go on to explore whether the social world can be understood differently from people's natural standpoint. If, for example, from that standpoint, it appears that there is no domination, no discrimination, no oppression, phenomenologists are therefore not in the position to question this assumption. Bourdieu also criticised the phenomenological

30 He uses the 'illusion' metaphor amongst others in (2004).

31 In what follows, I present a partial reconstruction of Bourdieu's theory of theory (of social practice). I mainly rely on his (1977) and (1980/1990), the most important sources for these views. These are, however, not always worked out in detail. He does not explain, for example, how subjectivism and finalism are related to each other. My reconstruction tries to bring out Bourdieu's criticism of the conscious/unconscious distinction.

32 Or, as Bourdieu also phrases it: the 'native experience and the native representation of that experience' (1977:2).

33 For a critical discussion of Bourdieu's criticism of the phenomenological stance towards the social world: see Throop and Murphy (2002).

approach for lacking a reflection on the phenomenological position. What are the social conditions of its own possibility? What is ‘the social meaning of the practical *epochè* that is necessary in order to conceive the intention of understanding primary understanding’ (1980/1990:25–26)?³⁴

In the reflections he performs as a theorist of social scientific theory, Bourdieu is performing, despite his criticisms of phenomenologists, a phenomenological analysis, this time however amending what phenomenologists neglected to do: to include the issue of the social conditions that account for the illusion of naturalness.³⁵ Theoretical reflection cannot, in this way, uncover those conditions in specific empirical cases. In the end, this job still needs to be performed by social scientific practice. But theoretical reflection can help to articulate, in general terms at least, what needs to be uncovered: what I have so far called the ‘veil of prejudice’, i.e. the dissonance between, in Bourdieu’s phrases, practical knowledge (native knowledge, doxa, doxic experience) and quasi-theoretical knowledge (native discourse, opinions, heterodoxy, orthodoxy). People themselves present an explanation of their social practices in their articulated opinions. But these present a distorted presentation of their social practices. What remains outside of the articulated opinions (the universe of discourse), i.e. what people do not say, is what people take for granted in a double sense: ‘what goes without saying and what cannot be said for lack of an available discourse’ (1977:170). As knowledge of the dissonance is not reflected in the explanation people offer themselves, neither is it available to their practical, tacit knowledge which Bourdieu, for this reason, calls ‘learned ignorance’ (*docte ignorance*): ‘a mode of practical knowledge not comprising knowledge of its own principles’ (1977:19). One reason at least why the dissonance is interesting to social scientists is that it may function to conceal to people different kinds of domination which social scientists may uncover in their research. It indicates the potential Bourdieu thinks the social sciences have for providing social criticism.³⁶

The second stance is objectivism. It typifies the position that social scientists, in principle, take towards the object of their research. The objectivist approach consists in an epistemological break with the social world they study, which it has to perform ‘in order to constitute the social world as a system of objective relations independent of individual consciousnesses and wills’ (1977:4). It is, therefore, a break between the theoretical knowledge of social scientists and the practical knowledge of people in the social world. Because of this break, which makes possible the employment of social-scientific research methods (statistics, for example), social scientists have access to a

34 Also see Bourdieu (1977:233, note 15).

35 Bourdieu criticises phenomenologists for forgetting to carry out an ultimate ‘reduction’ (see 1977:233, note 15), which suggests that he does not reject the phenomenological approach as such.

36 Bourdieu: ‘The dominated classes have an interest in pushing back the limits of doxa and exposing the arbitrariness of the taken for granted; the dominant classes have an interest in defending the integrity of doxa or, short of this, of establishing in its place the necessarily imperfect substitute, orthodoxy’ (1972/ 1977:169).

mode of knowledge that is not available to the phenomenological mode of knowledge: knowledge of ‘the objective relations (e.g. economic or linguistic) which structure practice and representations of practice’ (1972/1977:3).

Objectivism has to deal with other difficulties than subjectivism. Their methods allow social scientists to observe all kinds of regularities (apparent in, for example, price curves, income curves, employment rates, chances of access to higher education). These are not yet explanations of what goes on in the social world but are in the first place merely the regularities that show up in the scientifically performed observations social scientists made, for example, in the statistical or other data they collected (price curves, income curves, employment rates, chances of access to higher education).³⁷ The theoretical problem social scientists have to deal with is this: how to move from the regularities in one’s observations to an explanation of those regularities, particularly in terms of objective structures that can be ascribed to the social world? For that reason, they construct theories as models of reality. But the problem that arises here, according to Bourdieu, is that social scientists do not reflect on their own position as social scientists (as phenomenologists did not on *their* position). They are then in danger of conveying nothing more than their own standpoint by sliding from the model of reality to the reality of the model. It happens whenever social scientists are content with reifying abstractions: ‘the fallacy of treating the objects constructed by science, whether “culture”, “structures”, or “modes of production”, as realities endowed with a social efficacy, capable of acting as agents responsible for historical actions or as a power capable of constraining practices’ (1977:26–27). Social reality is then no longer viewed as an object of research but as a mere execution of the model.

Social scientists should instead aspire for a third-order knowledge, a theory of theory that provides insight into the *objective limits* of merely objectivist knowledge. They should reflect on the relationship between social practice and the models they construct to account for that social practice. Objectivism is therefore one-sided if social scientists only perform one break with the social world, the one that helps them to take distance from the social world by constituting social practice as an ‘*object of observation and analysis, a representation*’ (1977:2). They need to perform a second break by questioning:

‘the presuppositions inherent in the position of the ‘objective’ observer who, seeking to interpret practices, tends to bring into the object the principles of his relation to the object’ (1980/1990:27).

What the second break implies for the research that social scientists perform, according to Bourdieu, is that social scientists shift their focus from explaining regularities to

37 For Bourdieu himself, regularities were often the outcome of statistical research.

explaining the *production* of those regularities. And this means, as I understand Bourdieu, to ask for the possibility of ‘the coincidence of the objective structures and the internalized structures which provides the illusion of immediate understanding’ (1980/1990:26), i.e. for how objective structures are actualised (internalised) in people themselves, i.e. in their dispositions. This is the well-known structure/agency problem of sociology. It turns on the question of what it means that social structures ‘exist’ *independent* of individual consciousness and wills. Unless social scientists want to fall back into the reflex of reifying abstractions whose ontological status is unclear (‘transcendent entities’), they need a theoretical framework that articulates what Bourdieu calls ‘the paradoxes of objective meaning without subjective intention’, i.e. how objective structures somehow exist *in* people, unknown to them. Of course, this touches on the problem that interests us in this section: of understanding human practices as both social and hidden to people themselves (‘hidden’ of course, in a sense that yet needs to be explained).

The objectivist stance allows for two different answers: finalism and mechanism. Whereas finalist explanations reduce practice ‘to the conscious and deliberate intentions of their authors’ (in other places Bourdieu also calls this the intellectualist position), explanations of the second type ‘treat practice as a mechanical reaction, directly determined by the antecedent conditions’, (1972/1977:73).³⁸ As these present two ‘incompatible theories of practice’ (1977:27), they pose a dilemma for social scientists in choosing how to explain regularities. This dilemma constantly re-emerges in the social sciences. To solve the dilemma, social scientists have often taken recourse to the notions of ‘rule’ and ‘unconscious’, where ‘rule’ is equivalent to ‘conscious’. Historically the rule provided an alternative model to a crude, blind mechanism and made it possible of understanding regularity as a rule ‘that governs, directs or orients behaviour – which presupposes that it is known and recognized, and can therefore be stated – thereby succumbing to the most elementary form of legalism, that variety of finalism which is perhaps the most widespread of the spontaneous theories of practice and which consists in proceeding as if practices had as their principle conscious obedience to consciously devised and sanctioned rules’ (1980/1990:39). The model turns out to be a rehabilitation of subjectivism. This may sometimes function to sustain a critique that ‘naive humanism levels at scientific objectification in the name of “lived experience” and the rights of “subjectivity”’ (1977:4), as Sartre does. But this need not be the case. Rational choice theory is a social-scientific model that starts from, in Bourdieu’s terminology, subjectivist assumptions but does not necessarily share the concerns of naive humanism.³⁹ Introducing the ‘unconscious’ seemed to make it possible to find a

38 I think there is an interesting parallel between finalism and mechanism, as Bourdieu describes and rejects them, and regulism and regularism, as Brandom describes and rejects them. But I will not explore this here.

39 Bourdieu treats ‘rational choice theory’ as a form of voluntarism of which Sartre, in his view, is a representative. When he categorises ‘rational choice theory’ as a subjectivist, i.e. finalist, position, he implies that this theory views people as being bound, consciously, by rules.

middle-ground between old-fashioned mechanism and finalism. To illustrate this move, Bourdieu quotes Lévi-Strauss:

‘Modern sociologists and psychologists resolve such problems by appealing to the unconscious activity of the mind; but when Durkheim was writing, psychology and modern linguistics had not yet reached their main conclusions. This explains why Durkheim foundered in what he regarded as an irreducible antinomy (in itself a considerable progress over late nineteenth-century thought as exemplified by Spencer): the blindness of history and the finalism of consciousness. Between the two there is *of course the unconscious finality of the mind* [la finalité inconsciente de l’esprit] (1980/ 1990:40, Lévi-Strauss as quoted by Bourdieu, emphasis added by Bourdieu).

Relegating the obedience to the rule to the domain of the unconscious in this way seems to provide an escape from the dilemma between mechanism and finalism.

The conscious/unconscious distinction prevents social scientists from providing a proper explanation of the paradoxes of objective meaning without intention. We have just seen that Bourdieu rejects finalism, the position that understands agency in basically conscious terms because it implies a return to subjectivism, which he had already criticised. But he also rejects understanding ‘agency’ in terms of the *unconscious*. The move from the notion of the rule (‘conscious’) to the unconscious seems to solve the dilemma of mechanism and finalism, but the fallacy of:

‘converting regularity into a rule, thus presupposing a plan, is only apparently corrected in the hypothesis of the unconscious, held to be the only alternative to final causes as a means of explaining cultural phenomena presenting themselves as totalities endowed with structure and meaning’ (1977:120).⁴⁰

The turn to understanding finality in terms of the unconscious really presents a return to mechanism, although in a more refined variety. As it turns out, the conscious/unconscious distinction collapses into the distinction between finalism and mechanism.⁴¹ While finalism

40 Bourdieu rejects the rule as an explanatory device for yet another reason. The practice of social scientific theory has allowed the tacit merging of ‘rule’ and ‘unconscious’ in a way that the term ‘rule’ has been used to refer alternatively to a regularity, the social scientific model itself and finally to the norm that is consciously followed. This has allowed social scientists to shift between finalist explanations (rule as ‘ruling’, *règlement*) and mechanistic explanations (rule as ‘regulating’) without making a real choice (see, e.g. 1980/ 1990:37-39). Elsewhere Bourdieu calls the rule ‘the obstacle par excellence to the construction of an adequate theory of practice’ (1980/1990:103).

41 This may not be a perfect match because mechanism may be conceptualised without employing the conscious/unconscious vocabulary (as in behaviourism). But identifying them in the present context helps to understand Bourdieu’s criticism. Another way of putting it is that here I am concerned only with psychological mechanism.

presents the fallacy of conceiving people (as conscious) as the *origin* of their actions, mechanism is the fallacy of conceiving them (as unconscious) as *effects* (of antecedent conditions).⁴²

Against these positions, Bourdieu claims, as one of the insights of his reflections, that '[e]ach agent, wittingly or unwittingly, willy nilly, is a producer and reproducer of objective meaning' (1977:79).⁴³ Their experiences of the social world are structured 'without following the paths of either mechanical determination or adequate consciousness' (1980/1990:41) because they have been structured, however unknown to the agents themselves, to have a structuring force themselves. In being conditioned by social conditions, they themselves also structure the social practices in which they participate. The relation between structure and agency, objectification and embodiment, between *opus operatum* and *opus operandi*, between objective (social) structures and their internalisation in dispositions: this relation is understood by Bourdieu as a dialectical process. Finalism and mechanism, both predicated upon analysing people's psychology in the conscious/unconscious vocabulary, fail to recognise this. This is the essence of his objections against the unconscious as a theoretical tool:

'The hypnotic power of the notion of the unconscious has the effect of blotting out the question of the relationship between the practice-generating schemes and the representations – themselves more or less sanctioned by the collectivity – they give of their practice to themselves or others. It thereby discourages analysis of the theoretical or practical alterations that the various forms of discourse about practices impose on practice' (1977:203, note 40).

Following the strategy of the unconscious really reduces 'history to a 'process without a subject', simply replacing the 'creative subject' of subjectivism with an automaton driven by the dead laws of a history of nature' and it reduces 'historical agents to the role of 'supports' (*Träger*) of the structure and reduces their actions to mere epiphenomenal manifestations of the structure's own power to develop itself and to determine and overdetermine other structures' (1980/ 1990:41). The aim of the second break for objectivism, therefore, does not consist in choosing between finalism and mechanism, which only present a false dilemma, but of aspiring for 'a science of the *dialectical* relations between the objective structures to which the objectivist mode of knowledge gives access and the structured dispositions within which those structures are actualized and which tend to reproduce them' (1977:3).

42 Again Bourdieu mentions Sartre as giving the most exemplary view of people as being their own origins: 'Sartre makes each action a sort of unprecedented confrontation between the subject and the world' (1977:73).

43 From a theoretical, mainly formal perspective that is used to framing agency in terms of either/or and conscious/unconscious, this is a puzzling proposal.

6.4. Hidden prejudice as implicit intentionality

In this section, I will deal with what it means that prejudices are hidden, i.e. that people are ignorant about them. Previously (chapter 2, section 2), I have discussed the social-psychological notion of ‘implicit bias’ as one version of understanding ‘hidden prejudice’, namely by understanding ‘ignorance’ as being unconscious of the *content* of one’s prejudice. The challenge will be to propose a notion of ‘hidden’ that does not bring back in the conscious/unconscious distinction and which also fits within Brandom’s framework. In this section, I will propose to understand hidden prejudice as implicit prejudice. Here the adjective ‘implicit’, as Brandom uses it, should not be confused with the adjective ‘implicit’ that social psychologists use for their notion of implicit bias. I will define prejudice in Brandom’s terminology and see how it works as an *explicit* prejudice. Next, I explore a notion of hidden prejudice in terms of what the sociological Interpreter (or: sociological Scorekeeper) *attributes* to people with *implicit* prejudices. What will be crucial to this strategy is to carefully distinguish between what prejudiced people are (implicitly) committed to and what the sociological Interpreter is committed to. I will end this section by explaining how understanding hidden prejudice as implicit prejudice, from the perspective of the sociological Interpreter, is congenial to Bourdieu’s hermeneutic approach.

The obvious suggestion to understand prejudice within Brandom’s framework is to understand prejudice as an *incorrect inference*. This would be continuous with the Enlightenment approach to prejudice. But for understanding prejudice in a more consistent way with the use of this term in our western societies, a more detailed definition is necessary. Drawing on Brandom’s views on inferentialism, I propose the following definition:

Prejudice is a person’s commitment that certain, actual people *belong* to a certain social group on the basis of certain features (e.g. having a certain race, religion, culture, ethnicity, et cetera) that this person attributes to those people – in the sense that the attributed features are thought to be *intrinsic* features of those individuals, features therefore that those people cannot get rid of – while this commitment is used as an *entitlement* to the further commitment that a certain, *negative, practical* disposition (for example: not to work hard enough – ‘being lazy –, to steal whenever it is possible – ‘being criminal’) – one that is conceptually not necessarily linked with having an ethnicity, a religion, et cetera, as such – can *generally* be ascribed to people from this social group, while both these commitments are used as an entitlement (either for oneself, or for others, for example police officers) to treat people from this social group differently, in order to sanction these people whenever they give in to their disposition, or to prevent them from

giving in to this disposition.⁴⁴

One stereotype about African Americans in the United States is that they are more criminal. This seems to be confirmed by the fact that they are relatively overrepresented in American prisons.⁴⁵ What turns the stereotype into prejudice is when the statistical qualification ‘more criminal’ (in the sense of more likely to be imprisoned than people from other social groups) lapses into the commitment that African Americans generally have the *disposition* to be more criminal. This ‘lapse’ is possible because the underlying inference is one where the claim that people are African American is taken to justify the claim that African Americans have the disposition to be more criminal.

How a prejudice might work is explained by Brandom himself borrowing an example used by Dummett, who discusses the prejudice ‘boche’ and explains as follows: ‘The condition for applying the term to someone is that he is of German nationality; the consequences of its application are that he is barbarous and more prone to cruelty than other Europeans’ (Dummett 1973:454). Brandom treats it as a concept with material contents and analyses it as follows:

‘If one does not believe that the inference from German nationality to cruelty is a good one, one must eschew the concept or expression ‘Boche’. For one cannot deny that there are any Boche – that is just denying that anyone is German, which is patently false. One cannot admit that there are any Boche and deny that they are cruel – that is just attempting to take back with one claim what one has committed oneself to with another. One can only refuse to employ the concept, on the grounds that it embodies an inference one does not endorse’ (2000:69–70).

This is first and foremost an example of how an *explicit* prejudice works because, in this case, a prejudiced concept is linked to a certain term. Brandom also mentions for example ‘nigger’, ‘whore’ and ‘faggot’. The problem with such terms is that ‘they couple ‘descriptive’ circumstances of application to ‘evaluative’ consequences’ (2000:70).

But how would it work for implicit prejudices, i.e. prejudices that are not expressed? What Brandom’s discussion of explicit prejudices shows is that once prejudiced people

44 This definition does not imply that having prejudices can be ascribed only to people from relatively more powerful social groups. It aims at making explicit the inferential structure between certain (prejudiced) cognitive commitments and various practical commitments, regardless of whether people are in a position to act according to their practical commitments. Even if this were the case, the definition also allows the ascription of prejudices to people from relatively less powerful social groups because even they are capable, although in a more limited sense, of treating ‘others’ in a prejudiced way.

45 For a critical discussion of how African Americans are treated in America’s penal system and for what this implies for the (re)interpretation of this and other facts regarding African Americans: see Alexander (2010/ 2012).

express what they believe ('are committed to') about people from certain social groups, the expressions themselves have certain inferential consequences. Anyone who wants to avoid being called 'prejudiced' will have to choose her words carefully. This is different with implicit prejudices. As these are not expressed, one is less likely to be recognised as being prejudiced. Before moving on, I first need to explain the explicit/implicit distinction and how it is different from the conscious/unconscious distinction.

Brandom's normative pragmatism works 'with the general methodological principle that one should understand what is *implicit* in terms of how it can be made *explicit*' (2010:299). What is made explicit is a rationality that is already implicitly present in our actions. One possible candidate for understanding the hidden nature of some background beliefs, including at least what I call 'hidden prejudices', is, therefore, the notion of 'implicit'. The notion 'implicit' has been part of the philosophical tradition for a long time, even if it has not usually received philosophers' systematic, continuous attention. The term itself – in its Latin form as *implicitus* – was used in a philosophical sense, probably for the first time, in Boethius' translation of Aristotle's *Organon* and was later taken up by various medieval philosophers for further analysis.⁴⁶ In these medieval treatments, the 'implicit' was analysed as part of the phrase 'implicit proposition'. But even these early analyses show already, I think, how the meaning of 'implicit' easily shifts to 'non-propositional', the sense that is central to Brandom's notion of 'implicit'. Brandom, as we have seen, links the 'non-propositional' – the implicit, the non-articulated – to knowing-how.

Does Brandom's *explicit/implicit* distinction cover the same ground as the *conscious/unconscious* distinction? Both distinctions function in a different way if we understand them as *cognitive* conditions for people's beliefs and practices. After all, they are part of respectively a *pragmatist* and a *representational* theory of cognition. But perhaps they may still be taken to function in the same way as *psychological* conditions of people's beliefs and practices? Shotwell is one theorist who chooses the notion of 'implicit understanding' to understand practices of prejudice (Shotwell's own phrases: 'individual and institutional racism', 'systems of domination', 'gender formation'). For her discussion on 'implicit understanding', she draws, for example, on Polanyi's 'tacit knowledge', Bourdieu's 'habitus' and Gadamer's *Vorurteil*.⁴⁷ She is familiar with Brandom's *Making It Explicit*, but does not, somewhat surprisingly, make clear to what extent she takes over Brandom's account of 'implicit', a term that of course belongs to Brandom's technical vocabulary while neither Polanyi, Bourdieu, nor Gadamer uses it. Some of the senses of 'implicit understanding' are, however, consistent with Brandom's own use: implicit as non-propositional, as know-how ('skill-based knowledge'). But Shotwell also seems

46 See Giusberti (1982).

47 Gadamer uses *Vorurteil* ('prejudice') in a sense that we do not easily recognise today, namely as 'common sense', which for some Enlightenment philosophers represented all the incorrect beliefs (superstition, prejudice, et cetera) of ordinary people.

to treat ‘implicit’ as a substitute for ‘unconscious’.⁴⁸ I think this is a mistake. From the perspective of a representational (and therefore propositional) theory of cognition, the implicit in the sense of non-propositional implies a form of *not knowing*.⁴⁹ But this cannot be concluded from Brandom’s notion of ‘implicit’ and need not be concluded from other notions of implicit as knowing-how. In the next chapter (section 2), I will explore in more detail how, within Brandom’s framework, we may understand self-knowledge, including self-knowledge of one’s knowing-how. For now, I will assume that implicit commitments are still commitments that a person knows about.

My strategy to understand ‘hidden prejudice’ as implicit consists in making two moves:

1. construct ‘hidden prejudice’ as the commitments that the sociological Interpreter attributes to prejudiced people, but in a way that the attribution is as much as possible ‘descriptive’ rather than *critical*;
2. construct ‘hidden prejudice’ as consisting in a set of implicit cognitive and practical commitments which people *take for granted*, which here I will understand as: they do not recognise these commitments as *their* commitments, nor even as *commitments*.

It is, of course, possible for the sociological Interpreter to offer a reconstruction of the hidden prejudices that immediately reveals that they involve incorrect inferences. My ambition with making the first move, however, is to find out whether it is possible to describe commitments in such a way that the prejudiced person may recognise them as making explicit what she is committed to. The second move is part of an attempt to understand prejudices specifically as a social phenomenon. Whenever philosophers analyse beliefs in terms of ‘I believe that...’, it suggests that people have consciously, deliberately, committed themselves to a certain belief, in a sense that implies that they take full responsibility for it. If prejudice is analysed as a *social* phenomenon and, therefore, as part of my current strategy, of implicit intentionality, it implies, I think, a reduced sense of responsibility for prejudices. I will therefore put a few constraints on the commitments the sociological Interpreter might attribute to prejudiced people. She may not attribute a commitment in a form that suggests that the prejudiced person is committed, *from her own perspective*, to:

- a prejudice (because this is what the sociological Interpreter thinks about the commitment of the prejudiced person);

48 See, e.g.: ‘There are levels of implicitness, some moving into explicit consciousness, some becoming a background, and others occluded’ (Shotwell 2011:xviii).

49 See Marçal and Kisteamacher (2010).

- any commitment as an unjustified belief (again, because this is what the sociological Interpreter thinks about the commitment of the prejudiced person);
- any commitment as an individual commitment, expressed, for example, as ‘I know’ or ‘I believe’ (because then it would be an explicit commitment).

The challenge, therefore, is to keep carefully apart the commitments of the prejudiced person and those of the sociological Interpreter.

For my discussion, I will use the example of an employment recruiter in Sweden who, during the selection for job interviews, favours applicants with Swedish names over those with Arab names.⁵⁰ The sociological Interpreter might reconstruct the hidden prejudice as follows:

1. ‘I know that people with Arab names are less likely to be fluent in Swedish.’
2. ‘I am an employment recruiter, I am obliged to select only the best applicants.’
3. ‘I know that people with Arab names are less likely to be fluent in Swedish, I am obliged to select only the best applicants, so, I must favour applicants that have Swedish names.’
4. Conclusion: ‘I will select only the best applicants’.

This might represent the explicit inference to which the employment recruiter is committed. I will assume that the recruiter would subscribe to the conclusion and that, therefore, the fourth part of this inference is unproblematic. The first three parts, however, are problematic. I will start by analysing the first part.

The first part expresses a *cognitive commitment* for which the sociological Interpreter relies on the phrase ‘I believe’. Familiar in epistemology is the KK-principle: if you know some fact, you can know that you know that fact. While the issue of whether this principle is correct has been subjected to many detailed and technical analyses of its logic, my reason to doubt an ascription in the form of ‘I know’ is from a different order.

First of all, the choice between ‘I know’ or ‘I believe’ may already be problematic. The ‘I know’ suggests that the recruiter can provide reasons for his claim. But replacing ‘I know’ with ‘I believe’ would suggest that the recruiter may believe something else as well. The challenge here consists in keeping apart the perspective of the sociological Interpreter who keeps score of the commitments of prejudiced people, and the perspective of prejudiced people themselves. A scorekeeper may be justified in attributing a commitment to a person, saying, for example: ‘She believes *x*’. But from the perspective of the attributee herself, it may in some cases be incorrect to claim of herself: ‘I believe *x*’ or, for that matter, ‘I know *x*’, because she does not *believe (know) x*, she is *certain of x*!

Another disadvantage is that both phrases turn the commitment into an individual

50 Based on the research by Rooth (2007), already referred to in 1.2.

one rather than a social one. As I said before, I aim for a reconstruction of the commitment she does not recognise as a *commitment* or even as *her* commitment. Would it be possible to reconstruct the commitment in a different way, as something that would, from her perspective at least, reduce *her* responsibility for the commitment?

At this point, Wittgenstein's comments in his *On Certainty* (1969) may prove to be helpful, especially in the interpretation Moyal-Sharrock has given of them.⁵¹ In a sense, Wittgenstein also tries, as Bourdieu does, to understand what people take for granted. But different from Bourdieu, Wittgenstein is not concerned with social facts that are entangled with relations of power. He is seemingly unconcerned about distinguishing between natural facts and social facts that, for people, have the 'illusion of naturalness'. What is interesting is that, whereas Bourdieu seems to be focussed on unmasking prejudices as illusions, Wittgenstein's remarks can help us to understand what it means for people themselves to have commitments they take for granted. Hidden prejudices consist in what, from the perspective of the sociological Interpreter, are commitments prejudiced people undertake, but what prejudiced people themselves experience as certainties which, for that reason, they do not recognise as potentially requiring entitlement, not even when they are appropriately challenged.

Wittgenstein meditates on facts that belong to our basic, everyday experiences: being a human being, this hand being *my* hand, my parents being my parents. What puzzles him about such facts is that they seem to resist a standard analysis in terms of justified, true beliefs. People 'know' they have a body, and yet, under the pressure of philosophical doubt, they do not seem to be able to give proof of it. Wittgenstein leaves behind the standard philosophical strategy of analysing such facts and instead opts for making a distinction between certainty and knowing.⁵² What is interesting about his understanding of certainty is that Wittgenstein links it to the notion of mistake: '[...] when is something objectively certain? When a mistake is not possible' (1969, section 194). In her reading of this passage, Moyal-Sharrock explains that being objectively certain 'is not a matter of subjective or psychological conviction'.⁵³ What is a logical feature of people's certainties, in Wittgenstein's sense, is that doubting these certainties is 'logically impossible' (Moyal-Sharrock 2005:73). Wittgenstein does not exclude the possibility that a person may make a claim that seems to qualify as her 'mistakes' about what people usually consider to be a certainty. Normally people are certain 'that motor cars don't grow out of the earth' (1969, aphorism 279) and that their bodies are not, as Miss Natalija did believe (see chapter 1, section 4), subjected to ingenuously crafted influencing machines. But Wittgenstein denies that every false belief is for that reason a mistake: 'I should not call this a mistake, but rather a mental disturbance' (1969, aphorism 71).

51 Wittgenstein (1969); Moyal-Sharrock (2005).

52 Moyal-Sharrock (2005) interprets Wittgenstein as making this categorical distinction.

53 Wittgenstein's passage is more extensively quoted by Moyal-Sharrock.

To distinguish certainties from knowledge, Moyal-Sharrock introduces the metaphor of ‘hinge’, used by Wittgenstein himself in some of his explorations of certainty.⁵⁴ Moyal-Sharrock has explored the many examples Wittgenstein discusses and categorised them into various classes of ‘hinges’. While some ‘hinges’ are natural, others are either acquired or conditioned. Moyal-Sharrock explains:

‘When acquired [...] hinges are acquired the way we acquire skills or habits: through training or repeated exposure, not through propositional learning’ (2007:109).

What is important for understanding certainties, as opposed to knowledge, is that for people, these certainties are not open to doubt, not in any subjective or psychological sense, as Moyal-Sharrock warns us, but in the sense that it is ‘logically impossible’ (2007:73).

Understanding hidden prejudice as what for prejudiced people themselves are certainties, in line with Wittgenstein’s explorations and Moyal-Sharrock’s interpretation, provides a different way of making hidden prejudices explicit. It suggests that we look for a different way of attributing commitments to prejudiced people. The sociological Interpreter should acknowledge that prejudiced people take certain truths to be evident facts about the world and not commitments *they* undertake. What the sociological Interpreter attributes to them as a commitment they undertake, they themselves understand as a general fact, or in Wittgensteinian terms as a certainty, or in Brandomian terms as a fact for which they do not recognise any need for an entitlement because they do not recognise it as a commitment *they* undertake.

If the first part is rephrased, the inference attributed to the prejudiced recruiter now looks as follows:

1. ‘People with Arab names are less likely to be fluent in Swedish.’
2. ‘I am an employment recruiter, I am obliged to select only the best applicants.’
3. ‘People with Arab names are less likely to be fluent in Swedish, I am obliged to select only the best applicants, so, I must favour applicants that have Swedish names.’
4. Conclusion: ‘I will select only the best applicants’.

This still does not present the recruiter’s inference as hiding her prejudice. I will now turn to the second part.

One of the elements in the second part of the inference is that a *pro-attitude*, in

54 See, e.g. Wittgenstein (1969, aphorism 343): ‘We just *can’t* investigate everything, and for that reason, we are forced to rest content with assumption. If I want the door to turn, the hinges must stay put’.

this case, an obligation, is attributed to the prejudiced recruiter. In Brandom's technical vocabulary, this makes explicit a distinct *practical commitment*.⁵⁵ In the inference as reconstructed above, it is assumed that we need to make explicit (a) a pro-attitude in addition to the fact that the person involved is an employment recruiter and (b) yet another pro-attitude in addition to the previous one and the fact that 'people with Arab names are less likely to be fluent in Swedish'. The disadvantage would be that the need to assume a pro-attitude weakens the proposal to understand 'hidden prejudice' as 'implicit prejudice' because the pro-attitudes themselves already make it clear, for the recruiter herself as well, that she is making a difference between people with Swedish names and those that have not. Would it be possible, according to another explanatory strategy, to claim that 'I will select only the best applicants' may follow from both what is taken for a fact 'People with Arab names are less likely to be fluent in Swedish' and 'I am an employment recruiter'? For rational choice theorists, this is not a viable strategy as they assume that a decision ('I will select only the best applicants') is preceded by explicit deliberation. As we have seen, Bourdieu criticises them for not recognising that people's beliefs are constituted by social beliefs. But people, as Bourdieu suggests, suffer from amnesia of genesis. It creates for people the illusion of making personal choices whereas they conform to social dispositions of reasoning. Brandom also opposes, as does Bourdieu, 'rational choice theorists and others who approach the norms of rationality through decision theory or game theory' (2000a:30–31). His criticism, however, starts from a different position. Brandom is interested in how we may attribute to people when we see them performing a certain action, a certain practical reasoning to account for their action. What motivates Brandom's discussion is his concern for a proper understanding of our normative vocabulary and what it means to make explicit what people do or believe implicitly.

One way of explaining what people do, using a normative vocabulary, is to include into this explanation phrases that express normativity. Rational choice theorists understand normativity in terms of individual choice or, in more technical terminology, an individual evaluative pro-attitude. Standard phrases that express such a pro-attitude are, for example, wants, decisions, desires, preferences.⁵⁶ To explain Brandom's criticism of rational choice theorists, we will start with a simple example Brandom (2000a) himself discusses. When we observe someone opening her umbrella in a rainy street, we might think that, if we asked her why she opened her umbrella, the reason she would give is:

55 See Brandom (1993, Chapter 4, Section VI).

56 Brandom's discussion of the attribution of practical reasoning includes not only pro-attitudes in terms of personal evaluations but also in terms of more general evaluations. Of course, there are relevant differences between personal evaluations and more general evaluations to which an individual conforms. For my argument, it is enough that I discuss only the first. Brandom targets especially Davidson, but also Humeans. Nevertheless, I want to add some explanation here. The core idea of rational choice theory is that understanding actions in terms of reasons requires that we understand them in terms of a *conscious approval* of them. Bourdieu criticises this idea along the lines of the conscious/ non-conscious-distinction. Brandom's criticism of rational choice theorists may be interpreted along the lines of both a conscious/unconscious and an individual-social distinction.

‘It is raining’. This is, therefore, the practical reasoning we might ascribe to that person:

It is raining, so I shall open my umbrella.⁵⁷

This is still a *material* inference, one that does not yet use formal logic. We may put it into a formal scheme.

It is raining
Only opening my umbrella will keep me dry
⊢ I shall open my umbrella

What would change once we employ normative vocabulary in expressing the material inference? Rational choice theorists would treat the inference as incomplete: it is missing a premise, namely what that the person wants (or: desires, prefers). According to them, the inference should be represented like this:

It is raining
Only opening my umbrella will keep me dry
I want to stay dry
⊢ I shall open my umbrella

Simply ascribing to someone the belief that it is raining would not count as providing her with a good reason for opening her umbrella. We need to include in our ascription the attribution of a want (desire, preference) to stay dry.

Brandom does not in the first place disagree with attributing a decision (preference, desire) to people whom we see performing some action. He disagrees, however, with rational choice theorists about how to understand this attributed decision. It is one thing to add an extra premise, yet another thing to claim that this completes what would otherwise remain an incomplete and, therefore, *invalid* reasoning. To make his point clear Brandom contrasts two sorts of reasoning: material inferences versus formal inferences. The standard view on explanation is the formal one, which treats inferences as correct ones because of their form: an inference is correct if it is a formally valid one. The inference we have just seen, completed by rational choice theorists with an extra premise, is an example of such a formally valid inference. An example of a materially good inference may be this one:

It is raining
⊢ The streets are wet

57 This is one of three examples Brandom provides and which I will present here in a slightly different form (see Brandom 2000:84–89). While Brandom uses the ‘∴’ sign to indicate the conclusion, I use here the ‘⊢’ sign.

A materially good inference is correct because of the content of its non-logical vocabulary. Brandom favours an order of explanation that treats materially good inferences as prior to formally valid inferences.⁵⁸ This has implications for how he views practical reasoning, as in the example of opening one's umbrella:

It is raining
 ⊢ I shall open my umbrella

As we have seen, rational choice theorists claim this inference fails if we do not add the missing premise of a want (desire, preference). Brandom claims that this need not be the case. The inference presented here, a *material* inference, is already a correct inference which we may treat as good in the same way that the previous example provides a good inference:

1. If it is raining, the streets are wet.
2. If it is raining, I open my umbrella.

Both are materially good inferences. In the second case, the material inference would fail only if someone had a desire *not* to stay dry, as had Don Lockwood, the character played by Gene Kelly in the classic movie scene from *Singin' in the Rain*.⁵⁹ As I already suggested, Brandom does not disagree with attributing 'wants' or 'desires'. But he does disagree, as I understand him, with the claim that we can infer from such ascriptions that people have such beliefs in an empirical sense. The relevance of such ascriptions is not psychological, but normative: we should understand such ascriptions 'to make explicit in assertible, propositional form the endorsement of a *pattern* of material practical inferences'. If we conclude from what we see that a man in the rain, contrary to the usual patterns of practical reasoning, has a desire *not* to stay dry, this does not necessarily imply that other people have a desire to stay dry if they conform to that usual pattern of practical reasoning.

What we can learn from Brandom's discussion is that we do not need to make explicit a reason for the employment recruiter to explain why she is selecting only the best applicants. The advantage is that it contributes to a better understanding of the recruiter's behaviour as implicit, i.e. from her own perspective. I, therefore, suggest that we cancel the reasoning in the original third part and replace it with the conclusion. The inference would now be as follows:

58 This claim is part of his theoretical programme to explain the second in terms of the first. I will not further pursue his arguments, but for some explanation of this theoretical programme: see chapter 5, section 6.

59 Brandom himself mentions this example.

1. 'People with Arab names are less likely to be fluent in Swedish.'
2. 'I am an employment recruiter.'
3. Conclusion: 'I have selected only the best applicants'.

If this inference is a plausible reconstruction of the recruiter's reasoning (her *material reasoning*, as Brandom would say), it may show how this reference for the recruiter herself reveals not any obvious prejudice but instead a pattern of reasoning that she takes to be self-evidential. Only once the sociological Interpreter would provide a critical reconstruction of this reference, by making explicit every tacit step in the inference, would this reveal that the inference is an incorrect one.

My proposal to understand hidden prejudices as implicit prejudices, as sketched above, agrees more or less with Bourdieu's approach to social practices. Prejudiced people are not simply ignorant about how they perceive, judge and act. What they are ignorant about is the (explicit) interpretation of their practices. While people have their own opinions (*doxa*) about what their practices mean, implicit in their practices is a meaning (the 'learned ignorance') that the sociological Interpreter can make explicit. Implicit prejudices can be understood as commitments without the need to understand these commitments in any voluntaristic, rational-choice kind of way. Claiming that people *authorise* or, in Brandom's terminology, are committed to certain prejudices, is a way of *making explicit* what people do. And so, in this way, the sociological Interpreter may attribute hidden prejudices to people. This does not necessarily imply that people, from their own perspective, understand their practices as implying discrimination originating in prejudices they have authorised. The structure of 'authorising' is implicit in these practices; making it explicit in a claim that these people authorise their prejudices, is what the sociological Interpreter is doing. That people have committed themselves to certain prejudiced beliefs remains hidden from themselves: on the one hand, because they have forgotten about how they acquired their beliefs (what Bourdieu calls 'amnesia of origin'), on the other hand, because it is concealed by their own explicit views about those prejudiced beliefs (in Bourdieu's terminology: *doxa*). As a result, they experience the 'truths' they are committed to as something natural. Making explicit the underlying structures of authorising is the hermeneutical task for the sociological Interpreter.⁶⁰

60 See Blaakman (2012), in which a similar phenomenon was explored, drawing on the writings of mostly Bourdieu and Gadamer. Here the phenomenon that injustices can remain hidden from people performing those injustices, was called here 'hermeneutic invisibility'.

6.5. Implicit intentionality as social intentionality

How can we make sense, within Brandom's framework, of the veil of prejudice? In the previous section, I have discussed one aspect: how we may understand that prejudices are in some way 'hidden' to people themselves. But prejudices are also *social*, for example, in the sense that people have their prejudices *in common* with other people in their society. What would this mean? Does it simply mean that people happen to have the same prejudice? Or does it also mean that they *each know* that they all have their prejudices in common? In this section, I will explore the wider problem of whether Brandom's framework allows the attribution to people of collective commitments. I start with a brief glance at how similar issues are analysed within a mentalist framework and to which problem these analyses give rise. I then turn to Brandom's normative-pragmatist framework. Brandom himself has not engaged in discussing collective commitments.⁶¹ I will therefore propose how we may use his notions of inference and scorekeeping in a way that may solve that problem. For this, I will introduce the notion of 'public performance'.

Undertaking a commitment may raise the question (if one is appropriately challenged) whether one is entitled to that commitment.⁶² Someone – I will call her 'Alice' – may claim: 'I have seen a car accident happening near this crossing'. Suppose Alice was interviewed by a journalist and that she was telling how another bystander – Bertrand, someone she only knows vaguely, a distant acquaintance who happened to be there – had also witnessed the accident. Even if she might not have spoken to him, she might still claim: 'We have seen a car accident happening near this crossing'. Why is this person entitled to move from an I-perspective to adopting a we-perspective in making her claim or, as we may describe it, from making an *I-commitment* to a *we-commitment*? How will she justify her we-commitment, i.e. her claim that what she observed was observed by all bystanders, including the observation that they observed each other, observing the car accident? This is the problem of justifying collective commitments.

This problem is discussed here as a problem of *common* prejudices. But it is a familiar problem that has been discussed in philosophy in various versions, for example as the problem of *collective action/intentionality* or the problem of *common knowledge*.⁶³ Although both debates focus on different kinds of commitments, the first on practical commitments, the second on doxastic commitments, and therefore use somewhat different terminology, they both aim for analysing *collective* commitments. How collective commitments are defined and analysed as a problem has also basically

61 For a similar observation on Brandom and also Bourdieu: see Tuomela (2002:6).

62 See the discussion in this chapter, section 2.

63 For a general discussion of the problem of collective intentionality: see Schmid and Schweikard (2009a). For a general discussion of the problem on common knowledge: see Vanderschraaf and Sillari (2014). More recently, a debate is developing on 'joint attention', closely related to the other two debates. See, e.g. Eilan (2005); Peacocke (2005).

the same structure in both debates. Theorists from both debates mostly try to explain collective commitments in terms of individual commitments. Although such theorists differ in solutions they choose, for reasons of convenience, I will call them reductionists. To illustrate what in these debates the problem of collective commitments is supposed to be, I choose to home in briefly on the debate on common knowledge.

Common knowledge refers to a situation in which people not only know ϕ , but also know it together and may say ‘we know that ϕ ’. Mutual knowledge, however, simply means that some people who know ϕ , do not know this together and may not say ‘we know that ϕ ’, or, according to the formal definition by Vanderschraaf: ‘Informally, a proposition A is mutually known among a set of agents if each agent knows that A . Mutual knowledge by itself implies nothing about what, if any, knowledge anyone attributes to anyone else’ (2014:15). For example, if by some accident Peter knows that Richard knows the same secret, and if Peter confessed to someone else: ‘We (= Richard and me) know ϕ ’, this statement would be true, but only in the sense that ϕ is mutual knowledge. To claim that it is common knowledge for them would not be valid, for the simple reason that Richard doesn’t know he is included in this ‘we’. But what would justify a claim to *common* knowledge? Theorists on common knowledge usually reconstruct such a claim by starting with a claim, stated from a first-person singular perspective, and then working towards what may justify a claim, stated from a first-person plural. Their aim is often to explain we-commitments in a reductive way, to show how we-commitments can be made sense of in terms of I-commitments. To show how it works, we return to the example of Alice, who had witnessed a car accident. From what she observed – the car accident and Bertrand somewhere near – she might infer the following:

1. ‘I know that ϕ .’
2. ‘I know that Bertrand knows that ϕ .’
3. ‘I know that Bertrand knows that I know that ϕ .’

This sequence may be extended with many more steps. But at what point would Alice be justified to infer from the previous steps ‘We know that ϕ ’? While this (brief) sequence only provides a simple presentation of a (yet inadequate) solution, reductionists aim for reconstruction in a sense that individual commitments (in this case, individual *cognitive* commitments) in the end can add up to something that may count as a ‘collective commitment’ if only this sequence is supplemented with the appropriate conditions.

Both the debate on common knowledge and the debate on collective action/intentionality are structured by a distinction between individual and collective commitments – expressed in the case of *cognitive* commitments, by respectively ‘I know’ and ‘we know’. But the distinction, in how it is used, suggests a difference between two different *types* of commitments because they are contrasted in ways that make clear that they are not the

same and do not overlap.⁶⁴ In each case, it seems safe to assume that having an individual commitment by itself *excludes* having a collective commitment because, after all, it cannot by itself include what yet needs to be explained: having a collective commitment.

Can we get a firmer grip on how the notion of individual commitment is understood? If we home in, once again, on the debate on common knowledge, we see that *collective* cognitive commitment amounts to ‘common knowledge’. But to what kind of knowledge do *individual* cognitive commitments lead? The problem with expressions like ‘I know’ is that they are often ambiguous. If someone would say ‘I know where a treasure is hidden’, the phrase ‘I know’ can be used in at least four different ways. Below I will make these differences explicit by slightly adjusting the phrase ‘I know’.

‘Only I know φ ’: an expression indicating exclusive knowledge. A knows φ , and she is the only one who knows φ , or at least she attributes to herself exclusive knowledge of some fact φ .

‘I know φ too’: an expression indicating common knowledge. If A shares knowledge with B, he is allowed to say both ‘we know φ ’ and ‘I know φ ’. Or, to use another example: if a person is travelling with some friends to Paris by train, she may assert, depending on the context, both ‘I am going to Paris’ or ‘We (she and her friends) are going to Paris’. Which phrase she will use will depend on the context, i.e. on what her conversation partner will ask her.

‘Only I know that we know φ ’: an expression indicating a combination of mutual and exclusive knowledge. What is mutual knowledge is that both A and B know φ . But only A knows that φ is mutual knowledge for both.

‘At least I know φ ’: an expression indicating someone’s inconclusive knowledge as to which other people know too about φ : ‘At least I know that I know that φ , but I do not know whether other people, or which other people, also know that φ ’.

Even more uses might be added, but perhaps one of the above already matches how reductionists use *their* notion of individual commitments? At least it will not be ‘I know φ too’ because this meaning, of course, already presupposes what is to be explained: having common knowledge. Of the remaining three uses, the basic form seems to be: ‘Only I know φ ’.⁶⁵ Does this agree with the reductionist’s use of individual commitments?

64 See, e.g. Tuomela (2006), who uses for this distinction the terms ‘I-mode’ and ‘we-mode’.

65 The expression ‘Only I know that we know φ ’ can be reconstructed as a variety of ‘Only I know that φ ’ whereas ‘Only I know that φ ’ and ‘At least I know φ ’ have in common the self-attribution of exclusive knowledge, even if the second expression seems to express doubt about the exclusivity of having this knowledge.

Interestingly, reductionists do not seem to systematically analyse what they mean by individual commitments. Sometimes what indicates how they understand them is the expression 'private'. Barwise, for example, contrasts 'common knowledge' with 'ordinary (private) knowledge' (Barwise 1987:366).⁶⁶ It would mean that the reductionists' use of 'I know' (and similar expressions, such as 'I intend') comes close to what I identified above as 'Only I know ϕ '. In both uses, a person has certain unique knowledge in the sense that this *excludes* making a claim for common knowledge.

In the debate on collective action, Searle has developed an argument against the assumption that collective commitments should be understood as reducible to individual commitments. But the argument can be transferred to the debate on common knowledge. Paraphrasing the argument by Searle (1995 24-25), it may be claimed that an endless chain of 'I know' can never add up to a 'we know'. It would be interesting to see why exactly this argument holds. Applying the distinction between common knowledge and mutual knowledge will show that what the standard solution or its variants are able to explain is at the most that people have mutual knowledge. To distinguish the mutual 'we' from the common 'we', I will use the notation '*we*' to represent the mutual 'we' and 'we' simply to represent the common 'we'. (The mutual 'we' should not tempt us into thinking that it represents a group of people sharing common knowledge. It merely represents: 'I know that ϕ and I know that B knows ϕ , but B does not know that I know ϕ as well.')

If the sentences 'I know that ϕ ' and 'I know that Bertrand knows that ϕ ' are replaced by the sentence '*We* know that ϕ ', no meaning is lost, because at this point of the sequence, step (2a), from her perspective Bertrand still does not know that she too knows that ϕ . This reflection only occurs at the following step in the sequence, step (3a). But even reaching this step, mutual knowledge still does not change into common knowledge. Once again, the sentence 'I know that Bertrand knows that I know that ϕ ' also can be replaced by a sentence expressing '*we*' know: '*We* know that Bertrand knows that I know that ϕ '. The real problem is that no extra level of reflection will ever succeed in helping pass the point of mutual knowledge. The different stages (from stage (1) to infinity) show different levels of reflection. In this way, the reductionists are in danger of only recognising individual commitments. They appear to offer no correct notion of common knowledge as mutual knowledge is only the sum of consecutive individual commitments which have the same content in common. Their explanations recurrently end up in mutual knowledge and are not able to pass this point. No matter what the solution the reductionists offer, if they assume that collective commitments can be reduced to individual commitments, any solution will fail, and the most they are able to show is that a knower is able to understand mutual knowledge, which is different from common knowledge.

Brandom's normative pragmatics would seem to offer no better prospect for solving the problem of collective commitments. Brandom has a perspectival view

66 For other examples of 'private' that seems to point in a similar direction: see, e.g. Barwise (1987) and Tuomela (2006), Vanderschraaf and Sillari (2014).

on social intentionality, as we have seen (in section 2). He understands sociality to have an I-thou sociality. It means, for example, that people are scorekeepers: keeping score of other people's commitments as different from their own commitments. This has provoked criticism by Habermas (2000), characterising Brandom's account as *methodological individualism*.⁶⁷ What Brandom claims to be an I-thou structure is rather, according to Habermas, 'the relation between a first person who raises validity claims and a third person who attributes validity claims to the other person' (2000:345). Even more so, every new communication between people 'opens with an ascription that the interpreter undertakes from the observer's perspective of a third person' (id.). On this interpretation, Brandom's account would be in no better position to explain collective commitments. He would even seem to subscribe to a similar notion of commitment as expressing basically an individual perspective, while collective commitments strictly do not exist and should be understood in terms of individual commitments. I will argue that this is not the case. While the standard approach reduces collective commitments to individual commitments and Searle chooses the reverse approach, Brandom's normative pragmatism can be used, I think, for a strategy that differs from both. To outline this strategy, I will first return to the debates on collective commitments, attempt to uncover the basic assumptions and criticise them.

My claim is that Brandom's normative pragmatics dismisses talk of collective and individual commitments if this would imply a difference between *types* of commitments. Any commitment by itself is neither of an individual or a collective type. Instead, the distinction between adopting either an individual or collective perspective, between attributing to oneself either an individual or collective commitment, belongs to the conceptual content of a commitment. My proposal would be that, in contexts where it matters to assess who has a certain commitment and who has not, expressions such as 'I know' and 'we know' may function to express someone's insight into processes of knowledge acquisition. Here the basic distinction is between public and secretive, which is related to, but different from, the common/private distinction. If, as in the case of Alice and Bertrand, they understand that this situation – their presence near the car accident and their observation of the situation, including each other – is a *public performance*, at least for each other, they may infer from this that they have common knowledge about the situation. They are entitled to claiming: 'We know that ϕ '. In other cases, a person may infer from the situation that only she knows it. She has hidden her money and jewels in a little box and buried them in the ground, making sure that nobody saw what she did. If she would now claim 'Only I know where I have buried my box', it expresses her insight into what makes a situation a secretive one.⁶⁸

67 In contrast, Barber (2011) suggests a very different, favourable reading of Brandom's I-thou sociality, one that uncovers similarities between Brandom's and Levinas's analysis of I-thou sociality.

68 The point here is, of course, not whether she is entitled to that claim. It may be that someone saw, unknown to her, that she was burying the box. Her claim would still express her insight into what counts as a secretive situation.

The standard approaches to the problem of collective commitments privilege individual commitments. Only the phenomenon of common knowledge or, again more generally, of collective commitments is assumed to raise problems. Private knowledge is what represents 'ordinary knowledge' or, more generally, individual commitments are what represents ordinary commitments. My introduction of the public/secretive distinction may seem no improvement in this discussion. After all, what reductionists are trying to do is to make explicit what it means that a situation is a 'public performance'. If only they succeed in doing this, in terms of individual commitments, they will have explained collective commitments in terms of individual commitments. But I think their analysis fails at two points. First, they make the wrong choice in assuming that public performances require analysis. Second, their notion of individual commitment as expressing 'private knowledge' is fundamentally problematic: it differs from how I understand 'secretive knowledge. Even more so, adopting a notion of 'private knowledge' is likely to exclude, by definition, any collective commitment.

My claim would be that our *default* understanding of situations is that of public performances. It would imply that what we need to learn is not so much what makes situations into public performances but what makes them secretive. Young children, therefore, start with the broadest possible notion of 'publicness'. Gradually they learn the constraints on 'publicness'. As we grow up, we usually get better at recognising what turns a situation into less public. I think the idea of default publicness is consistent with Brandom's normative pragmatics. Undertaking commitments should be understood as being basically a public activity (either in doing or saying).⁶⁹ In their turn, the attributions of commitments should also be understood as being basically public. In Brandom's view, conversations do not usually function to make explicit what people agree on. What people agree on, their collective commitments, usually stay in the background unless any of the participants has reason to believe, and a reason to express, that a certain commitment may not be collective. Adopting a we-perspective in doing what we do is, therefore, the most common thing in everyday life, but we seldom go about making this explicit, i.e. to make assertions using the first-person plural. When people cooperate, they start from a background of common understanding.⁷⁰

The distinction between 'public' and 'secretive', as I use it here, differs radically from the distinction between 'common' and 'private'. Here it may be helpful to understand 'publicness' as a two-level notion: it stands in a contrastive relation to 'secretive' while

69 This aspect of Brandom's has received little discussion, as far as I know. As one of few Gibbard acknowledges this implication of Brandom's views and phrases what he takes to be Brandom's position as follows: 'We must reject the mode of explaining that has prevailed among philosophers: explaining public meanings as arising from interactions thinkers whose thoughts can be identified and understood independently of public practice' (2010:17).

70 Of course, we may still want to make explicit the conditionals that are involved in inferring from a certain situation that is a public performance. Brandom would claim, I think, that such inferences are material and do not allow for making explicit all conditionals since inferences on publicness will probably be non-monotonic. For an interesting discussion on monotonic inferences: see Brandom (2002, chapter 2).

it stands in a contradictory relation to 'private' (in its technical sense). It implies, for example, that 'secretive' does not mean non-public. When a person understands a situation as 'secretive', for herself and perhaps for still other people (if they have a secret meeting), it simply expresses her insight that the situation she (and her accomplices) is relatively less public than other, standard situations.⁷¹ But a secretive situation is still a 'public performance' in the sense that it is possible *in principle* that other people are able to acquire knowledge about this situation (perhaps a secret camera is registering their secret meeting). It means that my notion of 'secretive' differs from the notion of 'private' in the previous discussions. Although I suggested that 'Only I know ϕ ' may be equivalent to the 'I know' as reductionists use it to express 'private knowledge', closer inspection reveals that the phrases differ substantially. One context in which the phrase 'Only I know ϕ ' makes sense, from a Brandomian perspective, is the example of a person who all by herself, under secretive conditions, has hidden a treasure. But the examples that reductionists use in their debate are always those of public performance. It indicates a radically different notion of an individual commitment which, in their sense, is exemplary of ordinary knowledge. For example, their use of the expression 'private', as in 'private knowledge', suggests a link with the familiar notion of 'private language'. In that case, we should understand 'having individual commitments' as analogous to 'having a private language'. If that is correct, it would explain *why* reductionists analyse collective commitments the way they do. Individual commitments should be understood, on the 'private language' interpretation, in terms of mental states an individual has. The individual who introspects these mental states understands their contents as reflecting his first-person singular perspective, but in a sense that, by definition, stands in a contradictory relation to 'public', to 'we'. The individual's grasp of the contents of his own mental states is, in this sense, solipsistic. For an individual who attends a public performance and tries to infer what might justify his conclusion that the situation is common knowledge for him and the other people in the audience, therefore present a riddle if he cannot retreat to background commitments that are collective in a Brandomian sense.

People, as participants in a community, have been socialised into acquiring certain patterns of reasoning. However, these patterns may, from the perspective of the sociological Interpreter, present incorrect inferences. The process of acquiring these patterns was the result of exposure to public, iterative, non-propositional 'performances'. These patterns of reasoning are therefore public. In this sense, people have collective commitments. But as people take these patterns of reasoning for granted, understanding them as certainties, as natural, people do not recognise them as *their* commitments.

71 Admittedly, the term 'secretive' does not sufficiently cover what I try to capture here because this term implies that people have an interest in *making* a situation 'secretive'. This implication is not intended in my notion of 'secretive'. Probably 'private', in its ordinary use, is a better candidate for my notion as it also does not imply the denial of 'public performance', although it implies a smaller 'public' attending the performance. Obviously, in my current discussion, using the term 'private' for a different notion would lead to terminological confusion.

6.6. Conclusion

This chapter has shown how the pragmatist framework can be used to make sense of the veil of prejudice without taking recourse to the mentalist vocabulary of the conscious/unconscious distinction. It required explaining what it implies for understanding prejudices as a social phenomenon and as hidden from people themselves. I picked up on two ideas from Brandom's theoretical views: first, that the practices of rationality are intrinsically social practices (as the mutual attribution and scorekeeping of commitments) and, second, that rationality is explicit in what we say and implicit in what we do. Each of these notions, by themselves, did not suffice to explain hidden prejudice. My strategy was to combine them in the sense that the veil of prejudice is explained in terms of implicit social practices. To avoid an explanation of 'implicit' in terms of 'unconscious', I briefly explored Bourdieu's meta-theoretical considerations for his notion of 'habitus'. I then explored an explanation of hidden prejudices as incorrect inferences based on commitments that people do not recognise as *their* commitments, not even as *commitments*. To explain hidden prejudices as a social phenomenon, I chose to reinterpret Brandom's idea of mutual scorekeeping. I argued that we should distinguish – next to scorekeeping as the attribution to others of commitments that are different from mine – a more basic kind of scorekeeping consisting in the attribution of collective commitments to others they share with me. The first type takes place against the background of the second.



Chapter 7

A pragmatist notion
of responsible agency

7.1. Introduction

In the previous chapter, I have shown that the pragmatist framework can provide a very different interpretation of the veil of prejudice. If this is plausible, this framework might also provide a very different answer to my research question. This chapter aims to use the pragmatist framework for forging several notions to replace the liberal-mentalist notions of epistemic autonomy, freedom, responsibility, and manipulation. While presenting these notions, I will be gradually working towards a pragmatist answer to the research question: *What are the conditions for non-deliberative actions to count as holding responsible those people who perform practices of hidden prejudice?*

The following section will suggest an understanding of self-knowledge as *inferential* (instead of *introspective*). On this proposal, knowing about yourself, about the bodily movements you know as actions, about actions as socially relevant, is a matter of self-attribution. This proposal has two functions. It paves the way for explaining (section 3) how Brandom's account of the space of reasons as a social space may escape a different kind of determinism: social determinism. It also prepares the discussion on responsibility for practices of hidden prejudice (section 4) that self-knowledge and ignorance about one's prejudice are not mutually exclusive. On the contrary: the transition from ignorance to self-knowledge allows various shades. This insight has consequences for what strategies (including manipulative ones) one will pursue in holding people responsible for their prejudices. I will then suggest a pragmatist notion of non-deliberative action (section 5). Next, I return to the discussion on manipulation and show that non-deliberative action can be, but is not necessarily, manipulative. As this notion of non-deliberative action will still be rather abstract, I will suggest a strategy for adapting this notion to socio-political contexts by proposing an understanding of non-deliberative action as narrative performance (section 6) and illustrate this by discussing a case study (section 7). Finally, I will return to the issue of responsibility for hidden prejudices and criticise, using the pragmatist notion of responsible agency that I will have developed so far, the mentalist approach to this issue.

7.2. Epistemic autonomy revisited

Underlying the various notions that will be discussed in the next sections is some notion of epistemic autonomy. We have seen that it is part of liberal autonomy. We have also seen that epistemic autonomy can be assimilated, as *introspective* self-knowledge, into a mentalist framework. Is there any alternative? Does Brandom's normative pragmatics offer any notion of non-mentalist, i.e. *non-introspective*, epistemic autonomy? Brandom himself has not provided any account of such a notion. But I think his normative pragmatics does allow for developing an alternative notion of epistemic autonomy

as *inferential*.¹ Below I will not provide any fully detailed account, as this exceeds my purposes. But I will try to sketch what such an account might basically look like. I will start with resuming Brandom's account of people as rational beings, as discussed so far, and I will take up several notions to sketch a notion of inferential epistemic autonomy at two levels: first at a basic level as a notion of bodily self-knowledge (as proprioceptive knowledge), second at the more abstract level as a notion of social self-knowledge. My discussion of the second notion will be brief: in the next section, I will provide more comments.

People are claim-makers. In Brandom's view, it means that people's basic attitude towards the world is *normative*. As we have seen, having a normative attitude is an attitude people have for treating whatever they observe, whatever others do, whatever they themselves do, as either correct or incorrect. If we want to make explicit what functions as the criterion for treating something as (in)correct, we might add: according to their grasp of the world or, let's say, according to their concepts (or: their claims). Treating something as (in)correct according to a concept/claim may be done in different ways. Your friend's claim, for example, is that boletes are edible except for Satan's boletes (as they are poisonous). You might be strolling around through the woodlands with a friend, gathering edible mushrooms, and be warned by her that you have gathered the wrong mushroom, a Satan's bolete. Or your friend might gather herself an edible bolete instead. This presents two different ways of how anyone may 'act' according to her concept/claim on boletes. In one case, your friend treats your gathering a bolete as incorrect. In the other case, your friend treats her own gathering a bolete as correct. You, in your turn, have corrected your own concept of boletes (assuming you have one). While you always had thought that *all* boletes are edible, you now have learned from your friend's warning that you must distinguish between boletes that are edible and those that are not. (And next, you hope your friend will teach you how to tell the difference.)

The challenge is to understand how experience (in our example: the experience that you have gathered, unknowingly, a Satan's bolete) may be understood as the use of concepts. As Brandom sees it, one model would be that experience consists of two different kinds of activity: *applying* a concept and *altering* a concept. Your friend applies her concept of boletes as she gathers edible boletes by distinguishing them from non-edible boletes. Next, her warning helped you *alter* your concept of boletes as 'all edible' into 'all edible except for Satan's boletes'. On this model of experience, the *application* of a concept itself proceeds as a '*two-phase* story, according to which one sort of activity *institutes* conceptual norms, and then another sort of activity *applies* those concepts' (2002:213) or, in a slightly different version: '*first* one stipulates *meanings*, *then* experience

1 The account of self-knowledge that Richard Moran proposes, and which was discussed in chapter 2, section 5) may perhaps be characterised as a mixed view, i.e. as implying that self-knowledge is both introspective and inferential (via the therapist).

dictates which deployments of them yield true *theories*' (2002:214).² Different from this two-phase application is the *alteration* of a concept. Altering a concept consists of replacing one concept with another. As Brandom claims, this model works on the condition that concepts are thought of as complete or maximally determinate. Your concept of a bolete tells you how to distinguish a bolete from other mushrooms. Once your concept turns out to be incorrect (you mistakenly treated Satan's bolete as edible), you replace your old concept with a new concept of boletes (now including the idea that one type of boletes is not edible).

Brandom favours a different model of understanding experience as the use of concepts. It takes as a starting-point that concepts are determinate, but not in a complete or maximal sense. *Using* a concept is already both application *and* alteration: 'Conceptual content arises out of the processes of applying concepts – the *determinate* content of concepts is unintelligible apart from the *determination* of that content' (2002:215).³ It implies that the contents of our concepts have certain flexibility. They do not necessarily change drastically. We adapt them to every new case that we encounter and recognise as possibly relevant for that concept. The previous model starts with maximally determinate concepts and therefore has difficulties explaining where we get them from. Instead, the second model has the potential for explaining how people start their lives, as (very) young children, with concepts that have minimal content but learn to develop ever richer concepts as they get more experienced (on many levels of learning). I will not here provide such an account. For now, it suffices that learning concepts is (a) learning by doing and (b) constantly developing one's concepts (and this will usually mean: mastering ever more details, information, implications of any concept).

So far, I have sketched, however roughly and abstractly, how Brandom understands 'experiencing' as 'using concepts'. From this, I will try to extract a notion of epistemic autonomy in a basic sense: as *proprioception*, as knowing, for example, that whenever this arm is moving, it is *my* arm, and it moves because *I* want it to move.

First, I should like to suggest a constraint on bodily (agential) self-knowledge, a constraint that holds, I think, for any account that wants to make sense of this type of knowledge, both naturalistic and normative. Even if we have become adults and have become acquainted with our own body for a long time, even then, we only know *that* we are able to move our tongue, arms, legs (a form of knowing that consists only in 'doing'). We do not know, however, *how* we do this. One philosopher had expressed this insight already many centuries ago:

2 Brandom ascribes this model to Kant.

3 How using and altering one's concepts works according to this model is illustrated in the example of Druids living on a remote island having the concept 'is a bird'. Mark Wilson (1982) provides a rich and detailed analysis on how these Druids alter their concept (Wilson talks of 'predicate') of 'bird'. Also, see chapter 5, note 145.

‘When I want to speak, my tongue flaps about in my mouth; when I want to swim, my arms splash about; when I want to walk, my feet are flung forward. But I do not make that motion. I do not know how such a thing is brought about, and it would be impudent of me to say that I do what I do know how to do’ (Geulincx 1665/ 2006:33).

However well we become at knowing our bodily agency, however skilful we get in doing what we want to do, dancing elegantly, winning the Olympics, making a masterful drawing, we still would not know *how* we do it, in the sense that we cannot explain how a ‘thought’ or a ‘decision’ is able to ‘cause’ our legs and arms to move. This observation is consistent with a reliabilist approach that assumes that we have *reliable differential responsive dispositions* and assumes that we do not know the ‘mechanics’ of these dispositions. In Brandom’s account of normative creatures, it does not suffice that a person has reliable dispositions in relation to her own practical commitments. What holds for seeing ‘red things’ also holds for ‘dancing’. For her to enter the social space of reasons, she is required to understand how observational-practical.

According to the previously sketched model, my proposal is that we understand becoming acquainted with our body as the formation of a concept about our body. At the earliest stage, when we are only newly born infants, we will have only a rudimentary concept. Learning concepts is about learning *inferential* relations *between* concepts. Having one concept, therefore, requires having other, *many* other concepts. As we learn more and more concepts, we learn more about the world, and vice versa: we understand that our claim, if it is correct, allows making further claims, or, for our claim to be correct, it needs yet other claims to be correct. Even if infants may still be pre-linguistic or only linguistic in the most basic sense, their learning process might still be understood in terms of a (pre-)conceptual grasp of their body, of what sensation means what, of what sensation (‘my arm reaches for a toy’) follows from what (‘I want that toy’). This is a way of making sense of bodily knowledge as inferential: we acquire dispositions of knowing about our own bodies that assure us that we can do certain things.⁴

As I explained before, Brandom distinguishes between two capacities that people have: an *inferential* and a *non-inferential* reporting capacity.⁵ They are linked to two categories of concepts: ‘Purely theoretical concepts are those that themselves have only inferential circumstances and consequences of application. But these are linked inferentially to observational and practical concepts, and those links are essential to their contents’ (2005:434). Brandom concedes that our experiences, either what we observe (‘red things’) or what we do (‘movement of my fingers’), may by itself in many cases be *non-inferential*. Our ability to observe and to act depends on what was previously

4 The notion of epistemic autonomy at this basic level raises questions about what it means to have a self. For a discussion on this issue by Brandom: see his (2007).

5 See chapter 5, section 7.

called: our *reliable differential responsive disposition* (also: RDRD, see chapter 5, section 5). Our observations and our actions are therefore assumed to be reliable: for example, that I *see* ‘red things’ when there *are* ‘red things’, that my fingers *move* when I move my fingers. The *circumstances* under which we reliably apply the concepts of ‘red things’ or ‘movement of my fingers’ account for the non-inferential part of their conceptual contents. But for our observational and practical concepts, it is not enough that we have reliable experiences or that we *know* when to say, for example, ‘I see red things’ or ‘I move my fingers’. These concepts only enter the game of giving and asking for reasons once we also know how our observations and our actions fit in into the web of our other concepts: by committing ourselves to its inferential relations with other concepts, for example, by acknowledging that we did not see ‘green things’ when we claimed we saw ‘red things’. Once we tie our observational and practical concepts to inferential concepts with other concepts, they can function as what Brandom calls respectively ‘discursive *entry* transitions’ and ‘discursive *exit* transitions’.

Although I follow Brandom’s account on experience as using concepts, I will deviate from it on one point. Brandom distinguishes between two types of experiential concepts: observational and practical concepts. This distinction also returns in what was earlier discussed as Brandom’s Test-Operate-Test-Exit cycle (see chapter 5, section 7). He does not, however, comment on the origin of this distinction. Will a newly born infant ‘know’ that what is moving before her eyes are *her* arms? My answer would be a follows. At an early stage, infants do not know the difference between their bodily actions and the observation of bodily movements. What a baby observes as the arm *she is moving* (in exactly this ambiguous sense: it is her arm that is moving, but is it also she who moves the arm?) and what she observes through various senses: seeing, feeling (at various levels: her arm feeling resistance, feeling warmth, et cetera), does not by itself convey to her the message: ‘Hi, it’s me, your arm that you wanted to move!’⁶ During her intensive, evolving learning process, the young child will learn to understand the distinction between observing her fingers and making her fingers move, between seeing her arm waving and waving her arm. She will learn to attribute certain bodily movements to herself as what she did.⁷

This proposal for an inferential notion differs radically from the introspective notion of agential, bodily self-knowledge. We may see these differences in how both notions deal with cases of confabulation. Our RDRD’s usually function well, but sometimes they may fail. As a result, we will have experiences that are out of tune with ‘normal’

6 Anscombe (1957/ 1963, § 27) seems to be on this track when, in discussing how we know, for example, the position of our limbs, she remarks: ‘It may indeed be that it is because one has sensations that one knows this; but that does not mean that one knows it by identifying the sensations one has’ (1957/ 1963:49).

7 Even then, although it is possible to observe a kite flying in the sky without doing anything, the normal case will be that if you perform a bodily action, you also observe it, not in any consecutive sense, but in a sense that acting and observing fall it with each other.

experiences. As we cannot ‘introspect’ into the physical operations of our dispositions and therefore neither into their failures, we must ‘confabulate’, as psychologists call it, what it means. Earlier I discussed the experiment as performed by Wegner and Wheatley (1999).⁸ This provides a case in which people with normal RDRD’s nevertheless fail to notice that the movement of their hand was not *their* movement. Nisbett and Wilson, in their classic study (1977), conducted several experiments in which people failed to ‘introspect’ without noticing errors. Interestingly the failures of introspection are often understood as indicating some failure of rationality because people are understood to *confabulate* the reasons for what they did (or did not do) under experimental conditions. Of course, what is tested in such experiments is people’s self-knowledge according to a notion of *introspective* self-knowledge, a notion I reject.

The inferential notion deals differently with such cases. It analyses such cases according to people’s two capacities: their inferential and their non-inferential reporting capacity, the capacity to make inferences, and the capacity to reliably report on what one has observed or done. Brandom makes the following comment:

‘[H]ow do you tell whether something was an exercise of a non-inferential, observational reporting capacity, on the one hand, or whether it was the result of a process of inference on the other? It used to be that people just thought, well, ask them whether they made an inference; let them introspect and see. But the trouble is that it is possible to make all sorts of unconscious inferences. Any attorney who is cross-examining an eyewitness will find that the eyewitness herself is not very good at distinguishing between what she actually saw and what she inferred from what she saw, perhaps given common background knowledge.’ (Brandom 2008b:168).⁹

Although this was not a comment on experiments I just mentioned, I do think he would claim that such experiments are inconclusive. In the interpretation that Brandom suggests, the experiments show that people are ‘not very good at distinguishing’ between what they observe and what they merely infer. And therefore, when people are asked to justify a non-inferential report which they are certain about, they may retreat, in a split-second, to an inferential justification. Even if we leave this interpretation aside, the experiments only show that our RDRD’s sometimes fail in a way that we do not notice. But this is hardly news. Even if people had noticed their failing dispositions and had *chosen* to confabulate reasons, this only shows what I think is an essential feature

8 For an earlier discussion of confabulation: see chapter 4, section 5 (also see note 77 in that chapter).

9 One of the crucial scenes in the classic play *12 Angry Men* (Rose 1957) turns on this type of mistake: ‘[This witness] testified that in the midst of her tossing and turning she rolled over and looked casually out the window. The murder was taking place as she looked out, and the lights went out a split second later. She couldn’t have had time to put on her glasses. Now maybe she honestly thought she saw the boy kill his father. I say that she saw only a blur’.

of people's predicament as claim-makers: they have a normative attitude, an attitude to make claims about the world, even when confronted with anomalous situations that escape our usual claims, that seem to resist making justified claims. *If* a subject in the experiment of Wegner and Wheatley had noticed that her hand was moving even while *she* had not moved it, how should she have responded? She responded in a normal way: she ignored it.

Special cases of failing dispositions are when people suffer from 'asomatognosia', a certain loss of 'knowledge' about parts of the body. An extreme case is the Cotard-syndrome, when people suffer from total asomatognosia.¹⁰ People suffering from this syndrome usually claim that they are dead. Such cases are sometimes analysed as examples of failing rationality.¹¹ Here too, the inferential case provides a different interpretation. Cases of the Cotard-syndrome may be understood as indicating dispositions that fail structurally and radically. For that reason, people with this syndrome cannot ignore their experiences, even if the experts will know that these experiences are the result of failing RDRD's. As they are claim-makers, they cannot but try to make sense of their experience, one that to them may very well 'feel' like being dead. That is not a failure of rationality. It is a way of coping with a horrible, inescapable experience.

I will be brief about the notion of self-knowledge. While we grow old and 'master' our body, we also learn that the bodily movements we initiate are *actions*: they mean something to other people who begin to attribute all kinds of motives to us or teach us to behave according to certain social rules: 'Keep still on your chair', 'Don't pick your nose', 'Give a handshake to your uncle'. How do we know about our motives, about what we want, what we desire? The account of self-knowledge as introspective suggests that we look 'inside', that we 'learn' about our motives by becoming conscious of our mental states. I take it to be uncontroversial that people have 'inner speech'. People can 'rehearse' words and form sentences 'in their head'.¹² The question is how to characterise the significance of such 'inner speech' for assessing one's intentions. My proposal is that we understand our own motivations as based on self-attributions rather than on our interior intentions. Self-attributions of motives for actions we will yet perform, take the form of a *practical commitment* in Brandom's sense, i.e. as a claim that either we ourselves attribute to ourselves or which is attributed to us by other people. Self-attributions of motives for actions we have done in the past may be based on a mixture of the claims we remember we had for those actions and the claims that other people attribute to us.¹³ Other people, our relatives, our friends, may sometimes know us better than we ourselves.

10 See Dieguez and Annoni (2012).

11 See, e.g. Gerrans (2000).

12 For a philosophical discussion on 'inner speech' or, as Carruthers calls it, 'inner verbalisation': see Carruthers (1996). For psychological research on 'inner experience', including 'inner speech': see, e.g. Hurlburt (1990).

13 This resembles what Moran claims is therapeutic self-attribution. My claim, however, is that self-attribution based on what others (for Moran: the therapist) claim is less exceptional than Moran suggests. It is even the standard case.

The notion I propose of epistemic autonomy as inferential differs radically from a notion of epistemic autonomy as introspective. This can be traced to equally radical differences in their underlying frameworks, respectively the normative-pragmatic and the mentalist framework. First, the normative-pragmatic framework provides an *external* view, while the mentalist framework provides an *internal* view on intentionality.¹⁴ This difference does not turn on, for example, either denying or accepting the existence of thoughts as an ‘inner phenomenon’, but on how to characterise our thoughts. According to the mentalist view, we reflect on our thoughts as mental states. According to a normative-pragmatic view, we do not discover any mental states when we are thinking. When we are thinking, we discover meanings, but no vehicles contain these meanings. What we think about when we are thinking is about the world itself, the people in it, the situations in it, and about ourselves as being *in* that world. At a deeper level, the mentalist and normative-pragmatic views disagree about how to analyse intentionality. While the first view chooses to analyse intentionality *as mental states*, i.e. as having a vehicle (see the discussion on the *Vehicle Thesis*: chapter 2, section 6), the second view rejects this choice. Underlying these choices is yet a deeper-lying choice: while the mentalist view chooses to analyse intentionality as an empirical phenomenon, the normative-pragmatist view chooses to analyse it as a normative phenomenon.

Understanding the human mind as the source of action *in terms of knowledge of an interior* has been at the heart of many discussions in philosophy and sociological, economic, psychological and still other sciences.¹⁵ The dominant model for analysing the human mind is characterised by two theoretical decisions: taking the individual as the primary unit of action and analysing the ‘actions’ of the individual mind in terms of the vocabulary of consciousness. It has been an unfortunate development that the political individualism of contractarian-liberal theories has run parallel with the methodological individualism of philosophy of mind and of social psychology. Both types of individualism may have mutually reinforced themselves. While the central place of the individual has hardly been questioned in the philosophy of mind or social psychology, views on rationality have undergone a curious development. The reputation of what is most distinctive about the human mind – its rationality, its deliberations – is in one sense still untouched. The metaphors may have changed – shifting from rationality as ‘light’ towards rationality as ‘mechanism’, but Enlightenment theorists and philosophers of mind seem to recognise that the ‘mechanisms’ of rationality, although they may sometimes be defective, are largely reliable. But in another sense, the

14 Strictly speaking, from the perspective of Brandom’s normative pragmatics, there is no ground for making an internal/external distinction (in a cognitive, not a psychological sense). It is only implied once one assumes that acquiring knowledge about one’s own intentionality is a matter of introspecting one’s mental states.

15 The idea of an ‘interior’ is modern and began to dominate philosophical and literary works during the eighteenth century. See Lyons (1978); Davies (2008). In that period, the idea was also ridiculed: see, e.g. Toulmin (1979).

understanding of rationality has changed drastically. While rationality was understood during the Enlightenment as the ‘inner light’ or as what is performed by people themselves, its light or performances is nowadays understood as largely hidden from people’s sight: rationality has gone underground. It has shifted from the conscious to the unconscious realm, from introspection to the brain. This is the consequence of the introspection theory.

Various philosophers of mind and psychologists are nowadays committed to the introspection theory, paradoxically not because they want to claim that introspective self-knowledge really works – according to them, it does *not* –, but to claim that ordinary people themselves think that they have introspective self-knowledge and that they think it works. This paradoxical theory has remarkable implications. It would mean that people have no insight into their *real* motives or cognitions. Even while people *think* they know (e.g. about themselves, about their actions), they are really having a ‘delusion’. In cognitive terms, this is a radically different state than being in a state of ‘being deceived’. Rationality on the model of ‘being deceived’ still allows the possibility of recognising that one has been deceived. Rationality based on the model of ‘delusion’ does not offer this possibility. In this sense, they are not unlike schizophrenics having their schizophrenic fantasies.¹⁶ That various philosophers of mind and psychologists are committed to these implications is indicated for, example, by the consistent claim, as purportedly proven by various experiments, that people ‘confabulate’ (also see chapter 4, section 5; chapter 5, section 7), but also by claims that, for example, the ‘self’ is an illusion.

This brief, highly metaphorical diagnosis of what I think is wrong with the introspection theory may be rather abstract, rough or blunt. It is nevertheless an attempt to point at deep-seated assumptions that not only are debatable but that also impede the theoretical opportunities for exploring the potential that human beings have for changing their world, for the better, I hope. Rather than keeping to the notion of introspective self-knowledge, we should reject this notion as an incorrect one.

16 For a more detailed discussion: see Sass (1994). I interpret his discussions as a criticism of the mentalist framework. One of the implications of the mentalist framework, as Sass acknowledges, is to view people as locked up in their ‘interiors’, i.e. that they cannot rationally establish that there are really other people. This problem of solipsism is a recognised feature of the mentalist framework: see, e.g. Todd (1968). While schizophrenics have often been understood as living in their own world (in this sense solipsistic), Sass argues that this is, from an empirical point of view, not correct. He also continues to explore the philosophical implications of the false idea about the schizophrenic mind. The idea of introspective self-knowledge also has moral implications. See, e.g. Tauber (2005) who explores within a historical context how understanding the moral agent’s self-knowledge in internal, individualist-mentalist terms (‘reflexivity’, ‘introspection’, ‘self-consciousness’) is at risk of turning ethics into moral solipsism by neglecting how the self is also constituted by a community. A related analysis is offered by Yousef (2004).

7.3. Freedom revisited

In the previous chapters, I have outlined Brandom's normative pragmatics. In my view, it sustains a transcendental notion of freedom because it starts from a notion of rationality that *cannot* be understood in a naturalised, i.e. causal-mechanistic, way. In this section, I will explore how we may develop the *transcendental* notion of freedom into a *moral-political* notion of freedom. I will first deal with the possible threat of social determinism for Brandom's account. As a solution to this problem, I suggest that people's *general* ability to understand what is incompatible paves the way for the further ability to *deauthorise*. This also provides the starting point for forging a moral-political notion of freedom. My proposal will be to value the *ability to deauthorise* as a *freedom to disagree*, a moral-political freedom, therefore, that requires protection against possible interferences. Finally, I will suggest a few principles that might show how this freedom may be secured.

There is much that people will take for granted in their society in a way that sociologists sometimes call 'structure'.¹⁷ Bourdieu has attempted to understand such structures as embodied within people's practical attitudes through the notion of habitus. Earlier I suggested that we use Brandom's theory as a kind of sociological theory. Therefore, I proposed a notion of 'collective commitments' as *those* commitments that people within a community will attribute to each other as their common knowledge and which they do not recognise as either *their* commitments or even as *commitments*, because these commitments function as certainties (in a Wittgensteinian sense): they do not require any justification. Social structures could then be understood in terms of collective commitments. It is possible, I think, to understand 'collective commitments' on the model of 'habitus' because the views of Brandom and Bourdieu have several important similarities: rejecting the conscious/unconscious distinction to describe agency, resisting a mechanistic approach, adopting a pragmatic approach. Bourdieu, however, has spent more effort in explaining how people remain ignorant of a large part of their common (practical) knowledge. This has not been a concern of Brandom. If his theory is to function also as a sociological type of theory, it should incorporate a similar idea of people's ignorance about the social structures of their society, either as habitus or perhaps as a different sociological notion. The danger of putting too much emphasis on this ignorance is that the Brandomian sociological theory will give way to structures and power as the most important explanatory notions, but will lose sight of the normative dimensions of knowledge, more importantly of the potential that people have, as knowers, as claim-makers, to criticise these structures.¹⁸ In the worst case, a social theory will understand people as mere 'puppets' (as perhaps the sociological counterpart of the naturalistic 'automatons'). Bourdieu, for example, despite his intention to steer

17 Also, see section 6.3.

18 For a discussion of this dilemma for the social theorists: see, e.g. Bader (1988).

a middle ground between subjective and objective, between finalist and mechanist views of social practices, has been repeatedly charged for nevertheless sliding away into a socially deterministic view.¹⁹ Modelling ‘collective commitments’ on ‘habitus’ might therefore turn into a picture of people exercising their rational capacities, but only, unknown to them, within the limited space of contours that are determined by their social roles. In their turn, social roles may be determined by their ‘taken for granted’ collective commitments. It would mean that Brandom’s normative pragmatics, although it may show how people escape *natural* determinism, would yet, as a *social* theory, let them fall prey to *social* determinism.²⁰

Brandom’s theory can explain how people are not encapsulated by the deterministic forces of social structures. People’s rationality consists of various more specific abilities. One of them, as I have explained before (chapter 5, section 6), is people’s ability to recognise material (deontic) incompatibilities between commitments (see chapter 5, section 6). For the exercise of this ability, it is not relevant whether such commitments extend to the natural world or the social world. People do recognise, in actual societies, the many social incompatibilities in their societies even if historical or sociological researchers either do not choose or do not sufficiently succeed in bringing out their stories.²¹ The ability to distinguish material incompatibilities in social life provides the

19 See, e.g. Jenkins (1982) and King (2000). In an interview with Eagleton (Bourdieu and Eagleton 1992:114), Bourdieu even seems to confirm this determinism by saying: ‘Even in the most economic tradition that we know, namely Marxism, I think the capacity for resistance, as a capacity of consciousness, was overestimated. I fear that what I have to say is shocking for the self-confidence of intellectuals, especially for the more generous, left-wing intellectuals. I am seen as pessimistic, as discouraging the people and so on. But I think it is better to know the truth; and the fact is that when we see with our own eyes people living in poor conditions—such as existed, when I was a young scholar, among the local proletariat, the workers in factories – it is clear that they are prepared to accept much more than we would have believed’. Bourdieu’s writings, however, also suggest a way of escaping the charge of determinism by emphasising that people have a ‘sense of the game’ (*un sens du jeu*) and may be virtuosos as social agents. See King (2000).

20 Another way of understanding how Brandom’s normative pragmatics may slide into some form of social determinism is to ask how the I-thou structure really functions. What exactly is the model for their interaction? One option would be the educational model: the ‘I’ (parent/ teacher) that corrects the ‘thou’ (child/ student) in making mistakes. I assume that Brandom would recognise this model as one possible model. Another option would be the ‘punitive’ model: the ‘I’ sanctions the ‘thou’ for violating a norm. If these two would be the only two available models, then it would be unclear how Brandom’s normative pragmatics could escape social determinism. We would have to assume that a community somehow has certain first- and second-order rules/norms. But it would be unclear how they could change these rules in a sense that escapes social determinism. In this section, I suggest a third model according to which people may question all kinds of rules and norms and which some theorists refer to as an ‘agonistic’ model (see this chapter, note 27). This model expresses that the relationship between people should be understood, basically, as one of *ethical* intersubjectivity. A further issue, one that goes beyond the scope of my argument, would be how this type of intersubjectivity, which operates on a moral-political level, should be understood on a metaphysical level. Brandom argues that *epistemic* intersubjectivity is constitutive for the metaphysical level. While it might be claimed that epistemic intersubjectivity is more basic than ethical intersubjectivity, it also possible to argue for the reverse order: see Barber (2011, Chapter 4, Section 4).

21 On the methodological difficulties for researchers to make explicit the stories of the ‘vulnerable’: see Liamputtong (2007).

starting point for making a transition from a *transcendental* to a *moral-political* notion of freedom. Whenever people are subjected to discriminating practices that in whatever way disadvantages them, turning them perhaps into second-class citizens, they can develop a critical stance towards these practices. Recognising social practices as inconsistent is, however, but one step. It enables making yet another step: taking a stance towards the incompatibility between various (sets of) commitments. While people will always have (in Bourdieu's terminology) 'internalised', (in Brandom's terminology) 'committed' themselves (unknowingly) to or (in contractarian terminology) 'consented to' certain social structures that are disadvantageous to them, they may at some point withdraw their consent: they *deauthorise* the practices they previously 'consented' to. For deauthorising, it suffices that one denies the commitments that one previously endorsed. It does not need to include a concrete proposal for how to replace the discriminated practices with new practices.

Distinct from deauthorising is disagreeing: this is *expressing* – either in what is said or in what is done – that one deauthorises certain commitments. Disagreeing – expressing one's deauthorisation – has both transcendental and moral-political significance. Disagreeing affirms people's status as claim-makers in a moral-political (for example, Rawlsian) sense. It may challenge people's collective commitments in society. As such, it may contribute to public debate. But disagreeing also affirms people's status as claim-makers in a *transcendental* sense. First, to withdraw one's commitment to a certain claim, deauthorising it, should not tempt us to conclude that a person, for this specific commitment, ceases to be a claim-maker. On the contrary, deauthorising should be understood as undertaking a new commitment, even a particularly interesting one, namely a negative one.²² In specific social contexts, discriminated people are likely to develop their ability to deauthorise their previous commitments in their interaction with *other* discriminated people. Theoretically, however, it is possible to understand deauthorisation as a purely *intrapersonal* matter. In this sense, the ability to disagree is simply the ability to undertake a commitment that is inconsistent with a commitment to prevailing discriminatory practices. But deauthorising, as a merely 'interior' commitment, has no force unless it is expressed in what a person either does or says.²³ Only then, when a person expresses her deauthorisation, does it become an *interpersonal* matter. Only then is it possible for others to attribute to that person her disagreement with certain practices. And only then is it possible for others in their turn to agree or disagree. Deauthorising acquires its transcendental significance, therefore, only in disagreeing.

As discriminated people can develop a critical stance towards discriminatory practices, the actual exercise of their ability cannot in principle be interfered with by prejudiced people. Even expressing their deauthorisation may be done without inference if they express their disagreement in secret (micro-)practices of resistance. The

22 For making the point that a negation is not the opposite of an assertion: see Frege (1918); Geach (1965).

23 For this argument: see Anscombe (1957/ 1963, § 25).

experience of black women working for white families provides various examples. Those who worked as washerwomen not only washed the laundry of their employers but also of their own families. Or they might as domestic workers spit in the food, break some plates.²⁴ Prejudiced people *can*, however, interfere in these people's options for *expressing* their disagreement *openly*. This holds especially if the relations between *discriminated* and *discriminating* social groups have been asymmetrical for a long time – in the sense that discriminating social groups have historically been in more favourable circumstances in terms of having had more power and resources (in a broad sense: economic, cultural, political, et cetera). Under such conditions, it is possible, as Bourdieu argues (as we have seen in chapter 3, section 5), that social groups, both the discriminating and the discriminated, develop a social code of internalising disagreement with the social order. My proposal, therefore, is to value the *possibility* to disagree as a *freedom* to disagree.

I would like to contrast my proposal with the contractarian idea of freedom. For this, I want to highlight one of its features: *contractarian consent* as a precondition for the protection of people's freedom by social institutions.

Consent is a central notion to social contract theory. Previously we have seen (chapter 1, section 2) how early modern contractarian theorists had to struggle to find a proper notion of 'consent', one that expresses 'autonomy' without sliding into 'heteronomy' once people have authorised the social contract. I called it the *puzzle of responsibility as heteronomy*. Along a different line of explorations, yet another problem was put forward that is nicely summarised by O'Neill:

'One long-standing dispute is about the sort of consent needed for justification: should it be the *actual* consent of those involved or the *hypothetical* consent that would be given by beings with a distinctive (e.g. reasoned, informed, disinterested) view of the matter? If actual consent is needed, should it be explicit, or is tacit consent enough? Neither view of actual consent seems satisfactory for purposes of justification: we consent explicitly to too little, but (as it seems) tacitly to far too much. If, on the other hand, consent is only hypothetical, then it is quite obscure why it justifies. Why should the consent of hypothetical idealized rational agents, or of hypothetical beings in an ideal speech situation, or of persons in an artfully tailored hypothetical original position, justify the principles by which we are to live' (2000/ 2012:25–26).²⁵

²⁴ For these examples: see Kelley (1994, esp. chapter 1).

²⁵ Already during the early modern period, various philosophers began to raise such questions. See e.g. Hume (1772-1777a/ 1994). Although Vico is not usually mentioned in general discussions or historical overviews of social contract theory – he is not mentioned, for example, in Boucher and Kelly (1994) or in Cudd and Eftekhari (2018) – some Vico scholars think of him as a radical critic of the social contract theory. See, e.g. Lucente (1982) and Miller (1993:132) who summarises the view of Pompa, another Vico-scholar, as follows: 'Vico refuted the basic tenet of the social contract theorists, for in his view society could not rest upon a contract or agreement because contracts and agreements rest upon a promise, which was dependent upon an understanding of what a promise is, which in turn only comes about via social upbringing'.

Social contract theorists have given the notion of consent a central place because they assume that issues of justice can be best analysed as a transition – whether actual or hypothetical – from a situation *without consent* – as a state of nature – via an intermediate situation – an ‘original position’ (Rawls 1971) or an ‘initial bargaining position’ (Gauthier 1986:16) – towards a situation *with consent* when people agree on what each of them thinks are just institutions.

Although Brandom offers no political theory, his analysis of scorekeeping suggests quite the opposite order. This becomes particularly clear in an exchange of ideas and arguments Brandom had with Habermas over theories of meaning and how to construct them (also see chapter 6, section 2). For Habermas, the aim of communication consists in achieving agreement:

‘We understand a speech act when we know the conditions and consequences of the rationally motivated agreement that a speaker could achieve with this speech act. In short, to understand an expression is to know how to make use of it in order to reach understanding with someone about something’ (2000:346).

Brandom disagrees:

‘Conversational partners should not be pictured as marching in step, like soldiers on parade, but more as ballroom dancers, each making different movements (at any moment, one leads and the other follows, one moves forward and the other back, one sways left, the other right, and so on) and *thereby* sharing a dance that is constituted precisely by the coordination of their individually different movements’ (1994:363).

And the metaphor of dance recurs just one more time:

‘I have in mind thinking of conversation as somewhat like Fred Astaire and Ginger Rogers dancing: they are doing very different things – at least moving in different ways – but are coordinating, adjusting, and making up one dance. The dance is all they share, and it is not independent of or antecedent to what they are doing’ (Brandom in a letter to Habermas, dated 16th November 1997, quoted by Habermas, 2000:347).

Brandom’s point seems to be, as I understand him, that scorekeeping, communication in general and conversation in particular, already presupposes consent. But this consent, the commitments people share, cannot be specified as such, or, when it is specified, this can only be done if for each of the conversation partners it is clear from whose

perspective this consent is specified. Once one of the conversation partners makes a claim, from her perspective, about what is the consent between conversation partners, this also immediately opens the possibility for each of the other conversation partners to contest this claim, of course, against an unspoken background of what they agree about, until once again someone tries to make a claim about their consent.

What is interesting is that, in Brandom's view on the possibility of communication, the explanatory order that is central to contractarian theory (first no consent, then consent) is reversed into its opposite (first consent, then no consent). Some criticisms that were raised against social contract theory, of course, shared this idea: the state of nature already seemed, as an actual state, impossible because it denied what is needed to reach an agreement: having a common language. Once we transfer Brandom's view about the order of communication to the domain of political theory, it suggests that issues of justice can be best analysed as a transition from consent towards dissent, more accurately: from a situation in which consent is *assumed*, towards a situation in which dissent becomes apparent, or, once again, as the transition of a more or less (un)just society towards achieving a more just (less unjust) society.²⁶ The difference with a consent-based model is radical. The ambition of this model is to design, under ideal circumstances, what consent consists of. This should then function as the framework within which people may debate over their further disagreements. The dissent model accepts that circumstances are non-ideal: we are already in the middle of actual societies. This is our starting point. From here, we move further.²⁷

My suggestion is that we include the freedom to disagree in a notion of freedom of action, as a distinct, *cognitive* dimension of what it means to be free, and that we accord the freedom to disagree a central place in analysing how we can get from an unjust society to a less unjust society. I also suggest three principles that should be honoured, at a minimum, to safeguard the freedom to disagree: (1) having the opportunity to disagree; (2) being minimally burdened in using this opportunity; (3) having an equal balance of knowledge between social groups to which discriminated people belong, and other social groups. In general, the freedom to disagree is safeguarded if people can

26 I leave it to my readers how they prefer to characterise this transition.

27 For a related, more detailed discussion: see van den Brink (2005), who argues that the dissent-based model ('liberalism without agreement') provides a viable alternative to the consent-based model ('agreement-based liberalism'). It suggests that the dissent model is linked to *agonistic* notions of citizenship. See, e.g. Tully (1999); Norval (2007) and especially Fossen (2014), who connects Brandom's pragmatism with an agonistic model more extensively than I do here.

make an effective appeal to their right to disagree.²⁸ It means, in the first place, that people who want to express their disagreement have the actual opportunities to express their disagreement. It may mean, for example, that expressing one's disagreement is not sanctioned or disrupted. One way to hinder people in their freedom to disagree is to grant them the opportunity to disagree but make it difficult for them to use that opportunity. Voting is a way of expressing one's disagreement. In the past, at least African Americans suffered various obstacles that prevented them from effectively using their voting rights.²⁹ Finally, the type of knowledge that is at stake for the third principle is any kind of knowledge that may be relevant to the (re)distribution of power between social groups. The third principle requires that the distribution of this type of knowledge between people from various social groups is more or less balanced. It implies more general responsibilities for those social groups who relatively have more societal power (in various senses) than those who have not.³⁰ It may mean, for example, that they

-
- 28 The notion of the freedom to disagree is part of the more general notion of the freedom of political communication, which includes not only free speech in any ordinary sense but also expressive, non-verbal communication (see Frankenberg and Rödel 1981). It is intended to make explicit that 'discriminatory speech against vulnerable minorities should be treated differently from discriminatory speaking back and from discriminatory speech about less vulnerable minorities' (Bader 2013:8). For people from vulnerable minorities, the freedom of political communication, therefore, requires certain extra warrants. This is expressed in the first and second principles. They might function as part of a context-sensitive analysis of the freedom to disagree. The third principle, however, is different from the first and second principles. While these two highlight the practical aspect of freedom, in the sense of what people do, have a right to or are forbidden to do, the third principle brings out the epistemic aspect of freedom as what people know, should know or have a right to know. At this point, a comparison with the negative and positive notions of freedom is interesting. The negative notion of freedom is sensitive to the social context of actions, as non-interference, but is restricted to the practical conditions of action. The positive notion of freedom is sensitive to the cognitive dimension of freedom, as self-realisation, but neglects the social context of actions. The liberal-pragmatist notion of responsibility helps to highlight the cognitive dimension of freedom in social contexts: being free, in a socio-political sense, requires a certain kind of balance in having knowledge between various social groups within society if this knowledge is needed for having the opportunity to disagree with aspects of the socio-political order.
- 29 See Keyssar (2009), who mentions, for example, the following obstacles: literacy tests, 'understanding' tests, tax and property requirements.
- 30 Traditional notions of power tend to identify power with domination rather than with the power to resist (see Barbalet 1985; Sharp, Routledge, Philo and Paddison 2000a). However, I assume that, even in the most unjust societies, people from vulnerable social groups always have some form of capacity for civic action and, therefore, to resist other, less vulnerable social groups. Power is usually defined as some mode of the ability of doing, for example, in a non-relational sense ('power to') or in relation to either other people ('power over'; for a discussion of 'power to' and power over', see Göhler) or objects as the possible resources of power (see Bader 1991). What I call here 'balance of knowledge' between social groups indicates the importance not just of *having* certain knowledge but of having *access* to it. Therefore, it seems to indicate a different, cognitive dimension of power, i.e. as the ability of doing in relation to knowing, one which also differs from Lukes's classic analysis (1974) as highlighting the cognitive aspects of power. When less vulnerable social groups lack *access* to certain knowledge which is available to other social groups, for example, of how certain formal procedures work for expressing one's disagreement with certain decisions, this puts a constraint on their freedom to disagree. If other social groups have the resources at their disposal to give them this access, the principle of an equal balance of knowledge may offer a reason for providing this access.

should contribute, in whatever way, that people from other social groups also have access to good education.³¹ These three principles are abstract and should, of course, be analysed in more detail. For now, it suffices that they show how the freedom to disagree may roughly be protected.

7.4. Epistemic responsibility for social practices of silencing

In the previous chapter, I proposed a pragmatist notion of the veil of prejudice. However, it is of limited use for a *moral-political* account of practices of prejudice. While it offers an understanding of hidden prejudice in a non-mentalist and non-individualist sense, it still does not show in what sense prejudiced, discriminating people might be held responsible. Of course, they perform practices of discrimination, but they are still understood as being *ignorant* about their practices as prejudiced and discriminating. In the previous sections, I have outlined how we may understand discriminated people as escaping the challenge that is posed by understanding ‘social intentionality’ as a form of ignorance, a challenge that I described as the threat of ‘social determinism’. If in their turn discriminating people should still be understood as ignorant of their practices of prejudice, then we have not moved beyond the point that we already arrived at towards the ending of chapter 3. In this section, I will propose a notion of responsibility for hidden prejudices. This notion will turn on the question of whether people’s responsibility for hidden prejudices can be constructed – how? to what extent? – in terms of their knowledge/ignorance of their hidden prejudice. It is, therefore, a notion of *epistemic* responsibility. I will present two different, but complementary versions. In the first version, I will explain epistemic responsibility in an intercontent, *intrapersonal* sense. Next, I will move to a notion of epistemic responsibility in an *interpersonal* sense: epistemic responsibility for hidden prejudices is no longer described as an internal phenomenon, but as a social phenomenon, namely as a *practice of silencing*.

How does the pragmatist approach deal with issues of knowledge/ignorance for assessing epistemic responsibility? Again, as this approach does not use the conscious/unconscious vocabulary, it cannot adopt the strategy of explaining responsibility by modelling the knowledge/ignorance distinction on the conscious/unconscious distinction. So how would the pragmatist account deal with issues of knowledge/ignorance for assessing responsibility? It assumes that the self-knowledge that is central to epistemic responsibility is not a binary issue, as the conscious/unconscious vocabulary implies, but an *ambiguous* issue. I start by suggesting three different levels at which we may understand people’s self-knowledge as being ambiguous. At one level, we may understand knowledge in general, and self-knowledge more specifically, in various *modes*. Brandom’s explicit/implicit distinction, for example, suggests that people have knowledge

31 The Roma people, for example, have been denied in different ways good education. See Greenberg (2010).

in two or maybe three different ways: either explicit in what people say or think (in inner speech), implicit in what they do, or even a mixture of both. These two modes can be contrasted as knowing-that and knowing-how. However, I think we may distinguish even more modes of knowledge in a way that moves beyond Brandom's distinction but remains consistent with his account of inferential rationality. In the previous section, I have argued that people may understand the material deontic incompatibilities in social practices. We do not need to think of this 'understanding' as a Paulinian conversion or a 'decision' in the sense of rational choice theory. It may be more gradual. Neither do we need to think of it as explicit in what is said, or implicit in what is done. We might imagine how a person starts to grasp the social injustices done to her: in this sense, exercising her ability to understand what is deontically incompatible. However, she might not undertake any practical commitment and, for example, perform a secret act of resistance. She might not even make this grasp fully explicit in her 'inner speech'. Instead, I suggest that we recognise this pre-deliberative, pre-pragmatic understanding as a distinct mode of knowing that we may call *implicit understanding*. It would be interesting to explore this proposal in more detail, but for now, it suffices that I have made it plausible that (self-)knowledge may have different modes (even if we accept only Brandom's explicit/implicit distinction).³²

At another level, knowledge may be ambiguous because having knowledge about a concept (even within one modus) admits of gradation. Brandom sometimes talks about understanding as either *mastering* or *integrating* the inferential relations between concepts. Learning a concept is therefore learning new inferential consequences. But various factors may hinder people in the integration of their concepts, for example, because a reason for revising their concepts is absent, or because it is unclear how they may settle certain tensions between concepts. Below I will suggest various senses and degrees in which people may experience obstacles to mastering or integrating their concepts:

1. They do not know about the consequences of their prejudiced actions for people from social groups X-Z.
2. They know about the consequences, but they believe these consequences to be justified, in part by what they believe to be facts about people from social groups X-Z, and they do not know about counterevidence to these facts.
3. They know about counterevidence to these facts, but they do not know what is the turning-point for this counterevidence: when does it decisively show these facts to be false?

32 One interesting point that deserves more attention is the transition of one mode of knowing into another. My suspicion is that every such transition by itself activates the discovery of yet new inferential roles of a concept, simply because performing an act or articulating an idea introduces a person into a web of inferential relations that she may not have mastered.

Each of these suggestions, considered by itself, already allows for various gradations. For example, in the case of suggestion (1), when people buy clothes, they may not know about any of the consequences of their purchases for those who work in sweatshops.³³ But when they are white employers of African American domestic workers, they do have knowledge of some of the consequences of their actions for African Americans.³⁴

Knowledge may be ambiguous also because of an ability that I have discussed before (chapter 5, section 7) and that Brandom (2008:81) describes as ‘the capacity to ignore some factors one is capable of attending to’. For Brandom exercising this capacity is a normal part of people’s condition as knowers. As knowers and agents, people may follow many different paths to explore inferential relations between all kinds of concepts: ‘my left little finger and Bach’s second Brandenburg concert are not only different in countless ways, but are similar in that neither is a window-shade, nor a prime number, neither existed before 1600, and both can be damaged by the careless use of stringed instruments’ (Brandom 2008:81). But we will not usually devote our abilities to tracing in detail these kinds of similarities. Brandom does not provide any further explanation, but we may conjecture that people do not have the resources of rationality, time and energy to consider all possible relations between concepts. Instead, we will privilege (as the counterpart of ignoring) only some aspects of some concepts to understand their similarities and differences. The twin abilities to privilege and ignore certain aspects of concepts are constitutive of our ability to master inferential relations. What we privilege and what we ignore will be determined by the relevance of those concepts that concern us. While Brandom introduces the twin notion of privileging/ignoring for an account of human cognitive processes *in general* (he does not focus on any specific category of concepts), it should also be able to explain how people proceed epistemically in *social* contexts:

People know about counterevidence to facts they believe about people from social groups X-Z but ignore it.

It might be claimed that cases of *ignoring* are suspect. At the same time, as already suggested, people may not have the resources to consider and weigh what is presented as counterevidence.

The notion of epistemic responsibility that I have sketched so far, still leaves a lot of room for guessing what discriminating people ‘really’ know about their hidden prejudice. The three levels of having knowledge, including the various gradations of knowledge on each level, account for many ambiguities in our understanding of the knowledge/ignorance distinction. In each case, it is not immediately obvious what it implies for our attributions of knowledge/ignorance to people and therefore of their

33 For this example: see Iris Marion Young (2011, chapter 5).

34 For this example: see Kelley (1994).

epistemic responsibility. Therefore, I propose a second version of epistemic responsibility, one that is less open to ambiguity as it is understood in terms of certain practices: practices of silencing.

The earlier pragmatist notion of the veil of prejudice presupposes a situation in which, very much simplified, people from one social group perform practices of discrimination against people from another social group. Not only a sociological Interpreter, but also the discriminated people themselves can understand these practices by attributing to the discriminating people *commitments of prejudice*. But practices of silencing may license them also to attribute to discriminating people an *attitude of epistemic refusal*. To explain this, I will pick up again on some of Brandom's notions.

Earlier, we have seen that, in Brandom's analysis, reasoning as a social practice has a *default and challenge* structure. People within a community may attribute to each other entitlements *by default*. And they may also, for themselves, 'inherit' the entitlements they attribute to other people: whenever some people make specific claims ('commitments') and claim they are entitled to making this specific claim, other people may take over these specific claims. We might call this a 'default inheritance'. But this 'inheritance' of each other's claims and entitlements is valid if the original claim-makers can justify their specific claim whenever it is challenged by others, if they, to put it yet differently, not only take authority for their specific claim, but are also able to take responsibility for it.

The social practice of reasoning is basically an 'egalitarian practice' (Brandom 1994:241) in which, all participants are basically equals: for the social practice of reasoning to work, each claim-maker has in principle the opportunity of giving reasons to other claim-makers (taking responsibility their own claims) or asking others for reasons (holding them responsible for *their* claims). While Brandom tries to explore the logic underpinning the game of reasoning, he does not yet offer a sociological or political theory of reasoning: he does not, for example, engage in analyses of social disruptions of this game.³⁵ I will make a small step in transforming his theory into a sociological or political theory by introducing the notion of *socially structured default inheritance*.

My suggestion would be to understand the practice of 'default inheritance' as a socially structured practice: whether people will inherit (i.e. authorise) other people's commitments, will at least partly depend on their assessment of the *social* status of those other people. While each person may be described in many ways, any description that has consequences for how that person is valued or treated differently from other people, is a social status that is embedded within a web of inferential understandings of society. Relevant to my current analysis is that the attribution of a socially significant description, a specific social status (a certain ethnicity, for example) may have consequences for the way a person, because of the ascribed social status, is treated as a claim-maker. Default inheritances may be connected to certain social statuses in a positive or negative way:

35 He does pay some attention to practices that has the structure of superiors and subordinates.

positive, if people inherit by default certain commitments/ entitlements from people having certain social statuses; negative, if people *refuse* to inherit certain commitments/ entitlements from people having certain social statuses.

The default and challenge structure of social reasoning has the effect that many of people's collective commitments are part of their (unchallenged) background commitments and, in this sense, are taken for granted. Hidden prejudices, too, are part of a community's collective commitments (chapter 6, section 5). These remain in place as long as they are not challenged (chapter 6, section 2). Of course, they may be challenged. Reasons may be brought forward, by or on behalf of discriminated people, that question the commitments of prejudice held by discriminating people (which may also include discriminated people themselves!). Discriminating people may respond by taking these reasons into consideration and authorise them. Or they may take responsibility for their commitments of prejudice by providing counter-reasons. Any description of a community, or part of it, as performing practices of discrimination does not yet imply how practices of *challenging* commitments of prejudice will proceed. It would not be inconsistent, for example, to describe a community's practice as discriminating some of its members, while at the same time granting *every* member the same right to challenge anyone else's claims.

Practices of silencing suggest a different situation. While they presuppose practices of discrimination (not necessarily the other way around) and are also discriminating, they are analytically distinct from these. Once discriminating people are being challenged for their commitments of prejudice, they may engage in practices that have the effect of ignoring the challenging practices of discriminated people. In short: they silence them. Practices of silencing can be made sense of by describing them in terms of socially structured (positive and negative) default inheritance. Practices of silencing are therefore epistemic structures. They license that we attribute to discriminating people an *attitude of epistemic refusal*, expressing their claim that they are entitled, in response to reasons that challenge their commitments of prejudice, either to refuse to authorise these reasons or to refuse to being held responsible, i.e. to give reasons for their challenged commitments of prejudice.³⁶

Another way of understanding practices of silencing is to understand them as practices that persistently *prevent* discriminated people from effectively exercising their *freedom to disagree* with commitments of prejudice. This might seem like a problematic claim. If they are understood as a *response* to practices of 'giving reasons', then surely discriminated people have succeeded in exercising their freedom to disagree. But I think this argument is too hasty. At this point, it is important to understand in what sense the second version of epistemic responsibility is different from the first version. The attribution of an attitude of epistemic refusal should not obfuscate that this attribution

36 I think this 'attitude of epistemic refusal' might be integrated in, or is at least compatible with, Fricker's account of epistemic injustice (2007).

is licensed by a certain kind of practice that I conveniently call ‘silencing’ but which covers a wide range of very different, often non-verbal practices. If their commitments of prejudice are challenged, people may say: ‘What you are saying, is nonsense’.³⁷ But verbal forms of silencing may also take other forms:

‘Interviewers, recruiters, and job search committees carefully research the applicant’s references and look for any past history of “troublemaking.” During the interview, seemingly innocuous questions like “Why did you leave your last job?” or “Why do you want to work for us?” are sometimes slyly used to further probe a minority candidate’s potential for raising claims of discrimination. Through this process of screening, qualified minority candidates are denied jobs based on their potential for challenging racist prejudices and racist corporate cultures’ (Marvasti and McKinney 2007:71–72).

And people may also choose to simply ignore what their antagonists are saying.³⁸ Or they may even prevent their antagonists from challenging their commitments by persistently using violence any time against them. This suggests – although I will not further analyse this hypothesis – that effective practices of silencing should be analysed in relation to certain relations of power. In other words, the exercise of a freedom to disagree is not something a person can perform on her own. While many actions can be performed unilaterally, disagreeing is a social act. *Effectively* disagreeing requires other people to respond in the right way.

Practices of silencing indicate that discriminating people are not ignorant of their practices of discrimination, as they function to keep commitments of prejudice unchallenged and, therefore, to let practices of discrimination persist. From this second version of epistemic responsibility, it would follow that, while practices of discrimination might still be understood in terms of ignorance and therefore still leave room for a benign judgement about prejudiced people, practices of silencing exclude such an interpretation: a retreat to ignorance as an excuse is not possible.

The attribution of an attitude of epistemic refusal, which itself presupposes social structures of positive/negative default inheritance, provides a novel addition to standard definitions of prejudice. As I have remarked before, having commitments and entitlements is, in Brandom’s view, not only an intercontent, intrapersonal relation, but also an interpersonal relation. Standard definitions of prejudice are defined in terms of a person having faulty generalisations or stereotypes. They are analysed, therefore, in

37 Fricker (2007:9) provides an example of verbal silencing from Minghella’s screenplay of *The Talented Mr Ripley*.

38 Sartre’s account of bad faith (*mauvaise foi*) may be considered as an attempt to understand interpersonal situations of epistemic ignoring without invoking any notion of the unconscious (Sartre 1943). Also see the discussion by Waibel (2005).

terms of faulty intercontent relation between various notions one person holds (in that sense ‘intrapersonal’). What the attribution of an attitude of epistemic refusal may add to such definitions, is that it opens the possibility of understanding prejudices not just as a faulty epistemic relation between concepts held by one person, but also as a faulty epistemic relation between people holding different concepts. As practices of silencing have the effect of *ignoring* other people’s practices of challenging, they provide the interpersonal equivalent of what previously was discussed as the *intrapersonal ignoring* of counterevidence. When people present us with counterevidence to facts we hold dear, with reasons that challenge our convictions, we are forced to either privileging or ignoring. If we think it is relevant for concepts that concern us, for example, because we are committed to egalitarian ideals, we may privilege this counterevidence and explore its implications for our commitments. Or if we do not, we may ignore the counterevidence.

I have now argued for an understanding of people’s knowledge/ignorance of prejudices that is more ambiguous than the notion of implicit bias in a mentalist approach. While the veil of prejudice is a veil of ignorance for the mentalist notion of responsibility, for the pragmatist notion, it is a *veil of knowledge* as well, but a veil that allows some sight.

7.5. Non-deliberative action as non-manipulative influencing

In the preceding sections, I have argued that people can in principle be claimed to be responsible for their practices of hidden prejudice if these include practices of silencing. This fulfils at least one epistemic condition for holding people responsible for their practices of hidden prejudice. In this and the next sections, I will address the second epistemic condition of my research question, namely whether holding responsible may take the form of non-deliberative actions, or whether, instead, non-deliberatively influencing other people should be viewed as violating people’s epistemic autonomy, perhaps because it is manipulative. Earlier we have seen that lying and brainwashing illustrate two different types which, understood within the liberal-mentalist framework, violate *reflective* epistemic autonomy: the first was called reason-deceiving and the second reason-bypassing. While actions of the first type activate deliberative epistemic dispositions, actions of the second type activate non-deliberative epistemic dispositions (also see 2.4.). As the pragmatist framework starts from a different notion of rationality and a different notion of epistemic autonomy, it should have a different notion of manipulation.

In this section, I will explain, from within a pragmatist framework, what it means that people influence other people in a non-deliberative way and whether, or when, it would count as manipulative. Obviously, the category of ‘non-deliberative influencing’ is defined in opposition to deliberative influencing and therefore includes forms of *verbal* influencing that do not involve any deliberation in the usual sense, such as *storytelling*.

But I should mention here, as I indicated already in the general introduction, that I am interested in this category, especially for the *non-verbal* influencing it includes. So, my discussion below will mainly focus on non-verbal influencing as a subcategory of non-deliberative influencing.³⁹

The pragmatist framework takes *use* as more basic than *meaning* or, which amounts to the same, *practices* as more basic than *vocabularies*, what is *done* as more basic than what is *said* (also see 5.6.). In a world where people are no longer ‘proto-hominids’ but have evolved into beings that are able to make explicit what is implicit in doing, in a human world therefore where verbal acts – either spoken or written – have gained such prominence in shaping our understanding of the world, it may be a challenging task to claim priority for practices, in a non-deliberative, or even in a non-verbal sense. While in some social-scientific disciplines the importance of non-deliberative (non-verbal) actions has been recognised (as we will see in the next sections), it seems that in most social sciences and also in philosophical theories, human rationality is still exemplified in deliberative action, which really comes down to treating ‘speech’ as the centrepiece of human communication and human rationality, not because of ‘speech’ as a form of doing (i.e. as talking), but because of what it is said in the talking, and even more so, not just any talking, as in *storytelling*, but talking as deliberating, as setting up an argument. And yet, non-deliberative (non-verbal) action is the most normal thing in our human world. Non-deliberative actions can be as small and innocent as a smile or nudge, or also as small and discomforting as ‘stares, vocal inflections, hostile laughter’ (Ellison 1989:821). They may include keeping physical distance, but also touching (embracing, patting or slapping). In these various senses non-deliberative actions are the object of study of such disciplines as communication theory.⁴⁰ Non-deliberative actions may also include the use of objects, such as clothing.

Non-deliberative actions may be understood as a form of rational influencing, i.e. influencing that appeals to people’s rational capacities. According to a traditional view on rationality, non-deliberative action cannot be rational because, obviously, it does not involve deliberation: smiling, slapping or any other example of non-deliberative action is not usually perceived, by ordinary people, or analysed, by scientists and philosophers, as giving an argument in a conventionally recognisable way. According to a pragmatist view on rationality, however, they can be understood as a form of communicating commitments and entitlements. First of all, non-deliberative actions are *comprehensible* to other people, in the sense that they allow them, as ‘scorekeepers’, to attribute certain commitments to the agents.⁴¹ Second, as a further elaboration of the previous feature,

39 Admittedly the verbal/non-verbal distinction has been criticised by social scientists. For references and some discussion: see Payrató (2009). Using instead the deliberative/non-deliberative distinction may still not solve all ambiguities in identifying specific cases of cases, but it may perhaps be more distinctive.

40 See e.g. Knapp et al. (2014).

41 Non-deliberative actions function, as Brandom would say it, as a ‘discursive entry’ (2000:83). They enable observing people to make a ‘non-inferential report’ (2000:47).

those other people may use the attributed commitments as having an inferential structure: they may decide to treat the attributed claim as correct and accept what follows from it or what it assumes ('she smiled at me, so she must like me'). And finally, to understand non-deliberative action as influencing other people requires also that the performer of a non-deliberative action in her turn knows what commitments will be attributed to her based on her non-deliberative action and what the observers of her action will likely infer from it. These three features together show how non-deliberative actions may be understood as a form of rational influencing. The performer of a non-deliberative action may then be described as an influencer and the observer of that action as an influencee.

Once non-deliberative actions are described as influencing, it raises the question of how to distinguish, from a pragmatist perspective, between non-deliberative actions that are manipulative from those that are non-manipulative. For this we need to turn to the influencees and analyse what knowledge they have, or do not have, about the non-deliberative actions influencing them. As I commented earlier (in section 2.4.), manipulating other people requires that influencers have a certain knowledge of the methods or techniques to influence people. Of course, this holds for any influencing, manipulative or not. My proposal is that we distinguish between the general and the actual use of techniques of influencing. The first type refers to what people know of the *general* use of a specific technique of influencing, for example about how it generally works and in what contexts it is usually applied. The second type of knowledge refers to what people know of the *actual* use of a specific technique of influencing, i.e. whether they know *that*, and *when*, a specific technique of influencing is applied to *them*.

Using the distinction between the general and the actual use of a specific technique of influencing allows me to explain the difference between a mentalist and a pragmatist approach to manipulation. Mentalist analyses of manipulation, as we have seen (e.g. section 2.6.), have a fascination with introspective self-knowledge as a special kind of knowledge. Therefore, they usually assume that using a technique of influencing counts as manipulative if the influencee lacks knowledge of the actual use of this technique. A pragmatist analysis of manipulation is quite different. As the pragmatist framework understands all knowledge, including self-knowledge, as basically inferential (see section 7.2.), it leaves little room for analysing manipulation in terms of some special sort of self-knowledge. My proposal would be that, from a pragmatist perspective, using a technique of influencing counts as manipulative if the influencee lacks knowledge of the general use of this technique. As such, this claim is acceptable to both a pragmatist and a mentalist analysis of manipulation. What divides them is the interpretation of the case in which an influencee does have knowledge of the general use of a technique but lacks knowledge of its actual use. The mentalist analysis would treat this case as manipulative. But here, the pragmatist analysis, as I propose it, is less demanding: it would treat it as non-manipulative. Suppose that subliminal advertising has become a standard way of advertising in the cinema when movies are played and that this is

public knowledge for everyone in society: they have sufficient *general* knowledge about how subliminal advertising works and in what context they may reasonably expect to be subjected to it. It allows them to make a reasonable guess in what actual situations subliminal advertising is likely to be applied to them. In such cases, people in some sense know they are being manipulated. This will give them the opportunity to act on it: not go to the cinema, protest subliminal advertising or simply to resign in it.⁴²

Following my previous discussion, I now propose the following, pragmatist definition of manipulation:

Manipulative non-deliberative influencing is an influencer's knowing use of techniques that activate one or more of the influencee's non-deliberative epistemic dispositions and which result in the influencee adopting incorrect commitments which in their turn induce the influencee to perform certain actions, while the influencee has no (sufficient) general knowledge about these techniques.

As you may have noticed, this definition follows the structure of my earlier definition of manipulation, but with a few exceptions. First the class of manipulative actions is narrowed down to non-deliberative influencing as this is the focus of my current discussion (and therefore, 'lying' is excluded from this discussion). Second, the mentalist vocabulary has been replaced by a Brandomian terminology that should be understood according to an inferential notion of knowledge.

As I suggested before, theorists on manipulation have difficulties in drawing the appropriate boundaries to distinguish various kinds of manipulation. In a way, this is inherent to any attempt to come up with a taxonomy. One of the struggles for theorists on manipulation is to filter out what some have called 'benign manipulation'.⁴³ I think my definition provides a solution to some extent. Using perfume on a date might count as a form of benign manipulation. The proposed definition makes clear that we do not require to treat the use of perfume as a form of manipulation at all, if it is part of our public knowledge that people will use perfume when they go on a date. At the same time my definition is not intended to obviate all kinds of ambiguities. Analysing people's knowledge of their hidden prejudices admits of various degrees of knowledge, as we have seen. So, analysing their knowledge of the general use of techniques of influencing also admits of various degrees.

It might seem that the proposed definition already provides sufficient guidance

42 My point here is *not* that the self-attribution of knowledge ('I know I am being subjected to subliminal advertising') in such cases counts as knowledge in the strict sense that philosophers usually analyse 'justified true belief'. It may be the case that a person enters a cinema where people are not exposed, for whatever reason, to subliminal advertising. The point is that people in such cases are justified to surmise they will be exposed to subliminal advertising.

43 See e.g. Greenspan (2003); Todd (2013).

in assessing whether non-deliberative action is manipulative or not. My proposal is to add yet another ingredient: trustworthiness. Theorists on manipulation usually seem eager to explain manipulative action by analysing its impact in terms of an influencee's *intrapersonal* features (beliefs, desires, et cetera), features therefore that have no necessary reference to either the influencer or to the relation between the influencer and the influencee. In other words: mentalist analyses of manipulation tend to be non-relational. Instead, I propose that we move towards a relational approach to manipulation. Explaining manipulation should also include a reference to the relation between the influencer and the influencee in terms of trustworthiness: to what extent is the influencee justified or not in trusting the influencer?⁴⁴ The earlier definition may then be adjusted as follows:

Manipulative non-deliberative influencing is an influencer's knowing use of techniques that activate one or more of the influencee's non-deliberative epistemic dispositions and which result in the influencee adopting incorrect commitments which in their turn induce the influencee to perform certain actions, while the influencee has no (sufficient) general knowledge about these techniques *and has sufficient reason to trust the influencer for not applying these techniques to the influencee.*

This definition implies that manipulation cannot be defined independently of how people value one another in terms of trust. Adding to Brandom's theory of reasoning as a social reasoning, it would mean that people are also scorekeepers on each other's trustworthiness.

Despite the differences in understanding manipulation, both the mentalist and pragmatist approach seem to agree roughly on 'reason-deceiving' as a general characterisation of manipulation. One difference would be that 'seduction', which for a mentalist approach counts as a subtype of 'reason-bypassing', would from a pragmatist perspective also count as 'reason-deceiving'. But where does that leave that other subtype of 'reason-bypassing' within the pragmatist approach? So far, I have not included this subtype in my pragmatist analysis of my manipulation even though brainwashing, hypnosis and similar actions have been treated, in the previous chapters, as exemplary of manipulation and as essential to the mentalist analysis. First of all, the distinction between 'reason-deceiving' and 'reason-bypassing' makes sense only from within a mentalist approach, as it starts from a notion of introspective self-knowledge. Brainwashing for example counts as manipulation because it shows the failure of our introspective self-knowledge. The pragmatist approach that I propose starts from a

44 To my knowledge Brandom does not use the notion of trustworthiness, but it may easily be integrated into his pragmatist framework for which again I can draw on his distinction between the intercontent, intrapersonal and the interpersonal dimension of having commitments and entitlements.

different notion of knowledge, one that is inferential, social and interpersonal. It seems, therefore, that it cannot analyse, for example, brainwashing as manipulation in some other sense.

Theoretically the pragmatist approach can make sense of brainwashing and related forms of non-deliberative actions. Even such techniques as hypnosis and brainwashing can be understood as introducing commitments.⁴⁵ To understand what is special about such actions as manipulative, compared to 'normal' cases of manipulation, would be to assume that the manipulee is somehow prevented from integrating the new commitments within her original commitments. But this raises various questions that indicate a problem not so much with my pragmatic approach, but with our notions themselves of what we take to be brainwashing. Here I follow the argument that was already put forward by Dennett (1973). Apart from this philosophical argument, there is yet an empirical and a socio-cultural argument. Hypnosis, brainwashing or even subliminal persuasion have proven to be of great value for making great exciting stories in literature and moving pictures. But they should probably be viewed as mostly cultural or social fantasies.⁴⁶ So far, no experiment has made it plausible that hypnosis, brainwashing or even subliminal persuasion has the far-reaching impact that is usually attributed to it.⁴⁷ As these famous examples of manipulation have no philosophical or empirical plausibility, my suggestion, therefore, is that we stop talking about them as if they were plausible, even more so, that we stop using them as the material for our philosophical thought experiments. Thought experiments and metaphors derive their plausibility from the real world. But if nothing matches this type of manipulation, what plausibility remains for the thought experiments that make use of it?

The previous discussion now allows me to answer my research question from a pragmatist view: non-deliberative actions may count as a form of *holding responsible*, even those that from a mentalist perspective would count as manipulative (various forms of seduction) *if* people have sufficient general knowledge about the techniques of influencing involved in these non-deliberative actions.

45 If we did not make this assumption, we would have to assume that the hypnotised or brainwashed manipulee was acting like a zombie. But this does not seem to be implied in standard analyses of hypnosis or brainwashing.

46 More specifically on the cultural and ideological context of hypnosis: Andriopoulos (2000) and (2011). On brainwashing, particularly its cultural and ideological context: Carruthers, S.L. (1998); Melley (2008); Selisker (2016).

47 On subliminal persuasion: Pratkanis (1992) and Epley, Savitsky and Kacheliski (1999). On hypnosis: Orne (1972); Spiegel (1981); Rieber (2006); Murray (2016). Brain engineering is in full development for example for medical purposes. There seems to be no device so far that could function for mind control: see Foster (2006).

7.6. Non-deliberative action as narrative performance

In the previous section, I have concluded that, briefly summarised, non-deliberative actions may sometimes count as holding people responsible for their hidden prejudices without interfering with their epistemic autonomy. But the Brandomian-pragmatist notion of non-deliberative action is still too abstract and too far removed from the socio-political contexts of practices of hidden prejudice. It cannot therefore offer any ‘feel’ for what specific kinds of non-deliberative actions, within actual social contexts, might reduce practices of hidden prejudices. The mentalist framework, on the other hand, can provide such concrete proposals, such as administering an anti-prejudice pill or performing some type of brain-engineering, even if these proposals might be unsatisfactory and perhaps morally unjustified. I think my research question has burdened me with showing, however minimally, what it means, in a less abstract way and closer to actual socio-political contexts, to unravel the veil of prejudice and confront discrimination in a non-deliberative and non-manipulative way. In this section, I will therefore suggest a strategy for adapting this notion to socio-political contexts: it will consist in understanding *inferential rationality*, as what underpins non-deliberative actions, more specifically as *narrative rationality* and therefore to understand non-deliberative action as narrative performance. And it will be just that: suggesting a strategy. It will not provide a detailed theory as this would go beyond the purpose of my overall argument and, perhaps, probably, should be the object of a different research. My aim then will be modest: I will merely provide a preliminary and explorative account of integrating inferential and narrative rationality. It will result in the notion of *non-deliberative action as narrative performance*. This notion will be used in the next section to analyse some actual cases and explore whether non-deliberative actions might have the potential to unravel the veil of prejudice and confront discrimination.

Brandom uses a merely general idea of people as scorekeepers: he does not suggest that people may focus on certain types of scorekeeping, nor does he even distinguish between various types of scorekeeping. In contrast, my pragmatist account starts with the idea that we should understand people as having, in their attitude towards their life-world, a heightened sensitivity to any *appeals to agency* (either their own, or somebody else’s) or, to put in a Brandomian terminology: they are scorekeepers on social-political norms for agency. In socio-political contexts, this means that people will tend to focus on keeping scores on people’s agency, for example, their opportunities or their ambitions to do certain things.

The idea that people are scorekeepers on agency is rather meaningless if agency is understood just in reference to agency itself: as an isolated concept. Once agency is related to certain goals and actual situations, does it make sense to assess people’s agency. I assume that goals might be described in at least two distinct ways: as interests or as values. Following Fisher (1987), I will assume that the narrative view starts from the

premise that people are oriented towards goals as *values*. Values, whatever they are taken to be, present the ideals for the life-world. But they may turn up just to be ideals, as they fail to be realised in current, actual situations. At this point, values can be understood, in a normative-psychological sense, as a motivating force. When people authorise certain values, are committed to them, it means that they recognise the possibility of an *appeal to agency*: to take up the task of realising those values.

Whenever a gap occurs between an actual situation and a value, people will understand it as an appeal to agency. But this appeal might not be urgent, for example, because this value is viewed as less important than certain other values that people are working on. Or, if the value is viewed as important, people know that the responsibility for realising this value can be assigned to particular people or institutions who have the appropriate agency to realise that value. This might be called *role responsibility*. I assume that any society or community is familiar with certain *structures of responsible agency*. These represent what people, as a community, understand as the *distribution* of role responsibility, which answers questions as to who in what role, or which institution, ought to be responsible for realising which values. In whatever way we think of such structures as being embedded in a society, for example, as formalised in laws or protocols, it is important that we think of them as more or less effective to realise a society's values and as belonging to people's implicit social intentionality (section 6.5.). Whenever people perceive an appeal to agency, they will therefore not necessarily understand it as an appeal to *their* agency. Instead, they will understand it according to their society's structures of responsible agency, which means that first of all, they hold those agents responsible who have role responsibility.

The gap between a situation and an *important* value does pose a problem whenever it raises an *appeal to agency that cannot be met*. Structures of responsible agency tend to be ambiguous or flawed, often when a society is confronting complex or sudden, large-scale challenges, such as climate change or an epidemic. In such cases, people will perceive that appeals to agency cannot be met because agents with role responsibility either deny their role responsibility or because, while in principle affirming their role responsibility, they claim to be excused from their responsibility in this specific case, usually because they lack the ability (resources for example) to solve the situation. The first type of reason may indicate a crisis in a society's structures of responsible agency: if the agents having role responsibility deny their responsibility for this situation, then who has? The second type of reason suggests a crisis in the quality of a society's abilities, as when hospitals are no longer capable of providing proper health care as they are confronted with a large number of many patients with an epistemic disease.

Problems with structures of responsible agency can be solved either by reorganising the structure of how responsibility is distributed between roles or by reshaping the abilities of agency underlying structures of responsibility. Or they might simply be ignored. My suggestion is that, once such problems arise, people tend to suffer from

a lack of motivation either to understand or to solve them and that, therefore, their standard response will be one of ignoring them. This tendency might be understood as a psychological phenomenon. But for my account it is important to understand it as an *epistemic* phenomenon, i.e. as a problem of *epistemic agency*. This problem may be understood at a general level and at a more specific level. As I already indicated, my pragmatist account assumes that, in general, people have a heightened sensitivity to appeals to agency. But their understanding of agency and appeals to agency is embedded within their understanding of structures of responsible agency, which to a large extent is the result of socialisation and internalisation, processes which might be understood according to Bourdieu's notion of 'habit' or along the lines of my pragmatist notion of implicit, social intentionality. Therefore, when people confront problems with structures of responsibility, this exceeds, by definition, their 'habituated' epistemic agency. Instead, they will, in the absence of an understanding of a new, better structure, tend to think and act within the limits of their current understanding of structures of responsible agency and the roles they have within these structures. And so, they will tend to ignore the problems with structures of responsible agency.

It might be argued that this explanation is insufficient. As long as people ignore these problems, they must suffer from them. This by itself provides a source of motivation to try and solve them. To understand why people nevertheless lack motivation, we should turn to a more specific level of analysis: people's tendency to ignore these problems may be explained by their positions in society – the social group to which they belong, certain social characteristics they have, the place where they live, the jobs they have – and how their positions are related towards those problems with structures of responsible agency. It might be claimed, for example, that people, due to their social positions, will tend to lack motivation because they *do not personally suffer* from these problems or because they *do not even personally observe* them.

The previous analysis allows for the possibility that some people have the social positions to make a difference – i.e. to address problems with structures of responsible agency and change them –, but also lack the motivation of taking up that responsibility of making a difference. The next issue I will now turn to, is how these people may be addressed and motivated to take their responsibility. To understand how these people may be motivated – and again, in an epistemic, not necessarily a psychological, sense – into understanding and solving problems with structures of responsibility, I propose a *narrative view*. Adopting a narrative view on people means to understand people as understanding their life-world according to a *narrative structure*.⁴⁸ Their understanding

48 The 'life-world' is a technical term that has been used by a variety of philosophers, often working within a phenomenological tradition. For a discussion of this notion, especially in relation to notions of 'everyday life': see Sandywell (2004).

as having a narrative structure is what I call a ‘narrative’.⁴⁹ A narrative is a form of epistemic agency because it offers people an understanding of their responsible agency. But this epistemic agency should not be understood within a mentalist framework. The narrative understanding is not an ‘internal’, ‘mental’ epistemic structure. Instead, a narrative requires what I call ‘narrative performances’: a narrative performance will be any action having a narrative structure that is either explicit in what people *say* or implicit in what they *do*.

I will assume that, whenever people perceive a problem with responsible agency, they already try to make sense of it as a narrative. Whenever people perceive a problem with *structures* of responsible agency, this is different, as I already suggested: they will not necessarily make sense of it with any new narrative performance but will keep to their old narratives on responsibility. To this extent narrative performances have a *preservative function* as they preserve already existing structures of responsible agency. However, my hypothesis is that people may be motivated into understanding and solving problems with structures of responsibility if they understand them as a narrative, at least more easily than when they are offered merely technical arguments. To the extent that narrative performances are successful in this sense, they will have an *emancipatory function* as they will emancipate people from old structures of responsible agency.⁵⁰ First, I will further explain what I mean by ‘narrative’.

The twentieth century has witnessed, especially in its last decades, what has been called a ‘narrative turn’ (or: ‘narrative paradigm’).⁵¹ It provided new, interesting perspectives for doing research in a broad spectrum of disciplines, ranging from communication theory to literature, psychology and sociology. This required developing and specifying a narrative theory as a background for empirical research. Whatever the differences between the narrative theories that have been proposed and used in these various disciplines, the common starting point seems to be that people are viewed as storytellers and that this provides an important, perhaps even essential key to understanding a broad range of human actions, individual and collective, institutional and non-institutional. The implication for research is that what people do, is analysed as a form of storytelling. As storytelling requires an audience, it is also analysed as a form of communication between people.

For the narrative view that I will present here I will take my starting point from MacIntyre (1981, chapter 15) and Fisher (1987), who himself was partially inspired by MacIntyre. They both were discontent with certain dominant analyses of human action.

49 In formalist or structuralist theories of narrative a distinction is made between what is told (what happens, the events) and how it is told, expressed in terms of either ‘fabula’ versus ‘sjuzhet’ (Russian narrative theory), or ‘histoire’ versus ‘discours’ (French narrative theory) or ‘story’ versus ‘discourse’ (Anglo-American narrative theory). See McQuillan (2000) and Kernan (2010). For my discussion I will not treat this distinction as relevant.

50 For an analysis emphasising this function, especially for racial reform: see Delgado (1989).

51 See e.g. Czarniawska (2004; chapter 1); Hyvärinen (2006); Krizek (2017).

The first introduced the idea of people as storytellers in opposition to atomistic analyses that explain human actions in terms of ‘basic actions’. Fisher in his turn transferred MacIntyre’s idea into communication theory and used it to oppose what he called the ‘rational-world paradigm’ (1987, chapter 3), which analyses human action from a restricted notion of rationality that I take to be similar to what earlier in my argument has been termed ‘deliberative rationality’. Their own proposals are particularly useful, more than Brandom’s still highly abstract proposal, to understand human action in social contexts. At the same time, I find their accounts of what it means that people are storytellers, unsatisfactory, as they succeed insufficiently, I think, to move beyond the metaphor of ‘storytellers’ and explain how narrative performances may have an emancipatory function. Below I will briefly discuss some of their comments on the narrative view and to what extent narratives are different from, and similar to, stories and arguments. Next, I will propose, drawing on the notions of inferential rationality and the freedom to disagree, how we may understand narrative performances as non-deliberative and having an emancipatory function.

The idea of people understanding their own lives as storytellers might suggest that ‘story’ is transferred from the domain of fiction where it has its original place, to a domain, people’s non-fictional life-world, where this term is applicable only in a metaphorical sense. MacIntyre, however, seems to suggest a reversed order: ‘Narrative is not the work of poets, dramatists and novelists reflecting upon events which had no narrative order before one was imposed by the singer or the writer’ (1981/ 1985:211). In other words: people understand their life-world already as having a narrative order, even before their narratives were transformed, in the hand of dedicated, trained storytellers, into stories. In telling narratives people use, and have always used, some of the techniques of storytelling that we now consider to be the tools of poets, dramatists and novelists.

The idea also suggests that narratives are somehow *different* from arguments, for example, scientific, philosophical, bureaucratic or juridical ones. Especially Fisher is adamant on stressing the difference between narratives and arguments. While arguments express *deliberative* rationality, narrative and stories express *narrative* rationality. Explaining this difference, Fisher claims that ‘[t]he operative principle of narrative rationality is *identification* rather than deliberation’ (1987:66). For this identification to be successful Fisher distinguishes two criteria that should be met: probability and fidelity. As he explains these criteria, they seem to be largely relevant to both narratives and stories. ‘Probability’ for example refers to the coherence of the narrative/ story at various levels: structural, material and characterological coherence. Whenever an audience assesses the material coherence of a narrative/ story, it will compare and contrast it to other narratives/ stories they are familiar with. Central to both narratives and stories are also the characters: are they sufficiently consistent and coherent to be ‘probable’? A different criterion is fidelity which refers to the truthfulness of the narrative/ story.

These explanations still leave the question as to what extent narratives differ from

stories. It might be claimed that stories deal with the fictional while narratives deal with the non-fictional. Or it might be claimed that stories have a beginning, a middle and an end, while it might be debatable whether narratives also have an end.⁵² MacIntyre briefly discusses this distinction but rejects it.⁵³ Here I wish to emphasise another difference. Storytelling usually presupposes a basic dual model, according to which the story is told by a storyteller to an audience: the storyteller/ audience model. In the case of narratives, however, it is important to understand that this distinction is usually blurred: the audience itself is also active in shaping the story. Both MacIntyre and Fisher recognise that people may be authors of their narratives.⁵⁴ But I think their accounts do not yet sufficiently bring out why the storyteller/ audience model is problematic for understanding narratives. MacIntyre, for example, mainly discusses narratives as applicable to the relation that people have towards their own lives. But I am more broadly interested in the use of narratives for how people relate to large social and political issues. Fisher does show interest in such issues. However, as he is adamant about demarcating narrative rationality from deliberative rationality, he is in danger of insufficiently explaining the difference between narratives and stories. I think this happens when he, as we have seen, introduces the notion of ‘identification’. Without further explaining to what extent the identification of an audience with characters in narratives is different from the identification with characters in stories, he risks reintroducing the dual model of storyteller/ audience. The audience of a narrative would then be in much the same position as the audience of a book, a movie or a spoken tale: just reading, watching and listening. But while people may end their reading, listening or watching a story any time they want, it is not clear how, from the perspective of identification, how they can escape society’s narratives in which they are caught up. To this extent, Fisher’s notion of identification only helps to explain why narrative performances have a preserving function, i.e. contribute to maintaining the dominant narratives.

While MacIntyre’s and Fisher’s comments may still raise various questions on the similarities or differences between narratives, stories and arguments, for my account of the narrative view it is important to analyse narrative performance, first, as both rational and non-deliberative and, second, as having an emancipatory function. The first issue requires that we need to hone in on what notion of rationality underlies that of narratives. While Fisher explicitly chooses to talk about narrative *rationality*, as we have seen, he does not explore any notion of rationality that might underly and link both narrative rationality and deliberative rationality, perhaps because he is too much concerned with demarcating narrative rationality as a notion in its own right. At this point, I think, it would be fruitful to bring in Brandom’s notion of inferential rationality. Both Fisher and Brandom agree that a view of rationality in some deliberative, formal sense is too limited

52 For the view that stories have a beginning, middle and an end: see e.g. Krizek (2017).

53 MacIntyre (1981/ 1985:212).

54 See Fisher (1987:18).

and misrecognises a kind of rationality that is more basic than deliberative rationality. They also agree that this more basic rationality is really a communicative, social practice. Fisher is already content with calling this more basic notion 'narrative rationality'. I am not sure whether Brandom would agree. He might agree that narrative rationality is more basic than deliberative rationality in terms of how people develop their rationality. But he might claim that inferential rationality is still more basic than narrative rationality. I will choose this strategy and suggest that we accept his notion of inferential rationality as what narratives, stories and scientific arguments presuppose: while arguments consist of so many argumentative steps, each of which functions as an inferential antecedent to the following and as an inferential consequence to the previous, with each chain of argumentative steps pushing the overall argument further, stories and narratives consist of so many situations, each of which functions as an inferential antecedent, as what characters need to respond to, and as inferential consequences, as what results from earlier responses of characters, with each sequence of situations propelling the overall plot. Central to the inferential rationality that underpins narratives and stories are the commitments to values, which may be viewed as a specific type of claims. Narratives and stories are premised on the idea that commitments to values have inferential antecedents and inferential consequences. They have this idea in common with ethical theory. But while ethical theory will try to understand and formulate the basic rules that should underly people's actions when they are oriented towards values, narratives and stories can be viewed as *exercises* in understanding responsible agency: they explore what may happen when the values to which characters are committed, are somehow disrupted, for example, because they fail to provide guidance for action, because they are thwarted (by circumstances, fate, other characters), compete with other commitments. Narratives and stories explore, one might say, what people will do when they experience a problem with the inferential roles of their commitments to values.

So far, I have indicated how narrative performances may be understood as a kind of rational yet non-deliberative action. I also suggested how narrative performances have a preservative function. But we still have not seen how they may have an emancipatory function. A narrative performance may have an emancipatory function if it succeeds in releasing a person from a deadlock she has or experiences in her responsible agency. But how may we understand a narrative performance as emancipatory if, as was suggested previously, people identify with current, dominant narratives and if we cannot explain what it means that they can escape this identification? I will make three suggestions for explaining this.

My first suggestion is that people should be understood as having the ability to disagree with current narratives. This builds on the notion of epistemic notion I proposed earlier. Its significance in moral-political contexts is the freedom to disagree (see sections 7.2. and 7.3.). Whatever narratives dominate people's lives, people still have their own experiences. It would be misleading to call such experiences 'personal', because in a way

they are always personal, social and political at the same time. They can fulfil a critical role, because they provide people with the epistemic resources for tracing certain deontic incompatibilities within the dominant narratives. The idea of the freedom to disagree therefore suggests that dominant narratives in a society are not all-determining. People may identify with their society's dominant narratives, but it does not mean that they cannot escape them: they may disagree with them.

My second suggestion is that people should be understood as having the ability to move beyond merely disagreeing and to imagine alternative narrative performances. The idea of a narrative inferential rationality allows for the possibility of distinguishing between people's narrative ability – their ability to identify with the characters of any narrative/ story – and their actual narrative performances. It implies that no narrative, however it is constructed, can absorb or monopolise all other narratives. Narratives are open-ended accounts of responsible agency. Whatever narrative actions people perform in support of society's dominant narratives, this still leaves room for people to imagine alternative narrative performances. Their own experiences, but also those of others, again provide an important source, this time for experimenting with alternative narrative structures.

My third suggestion is that the identification within narratives implies a normative force that the identification within stories lacks. Identification within narratives is in many ways similar to identification within stories. When an audience identifies with the characters of either narratives or stories, it means that they understand to what values the characters are committed, but also what action should follow from those values and what sacrifice, benefit, pain or happiness will follow from performing that action. But identification within narratives is different from identification within stories in at least one important sense. When an audience identifies with the main characters, it will not only understand their values but also value the characters' values as *theirs*. This may mean that the audience will either adopt *new* values which it had not before or value certain values as *more* important as before. In both cases, the identification implies that the narrative has normative force. In Brandom's vocabulary: identification in narratives implies that the audience authorises the main characters' commitments to certain values. This is not necessary for identification within stories. We may identify with characters in stories even if we do not share their values or even if their values seem alien to us. If we do not share the values of a character in a story or do not even understand those values, we may nevertheless identify with this character in the sense that we are able to understand what a certain person, having such values, might do, might be afraid to do, might be unwilling to do, et cetera.

The imagination of alternative narrative performances, as either an adaptation of or a substitution for current narratives, may at first be limited to what people say or think. As such it fails to motivate them to act according to the values embedded in these narrative performances. Even then, the identification creates the potential for

motivating people. As long as people only imagine alternative narrative performances, perhaps secretly, and still perform narrative actions, publicly, according to current narratives, they must grapple with an epistemic dissonance between the dominant and the new, imagined narrative. Whenever this happens, this is the source of a 'tension', in a psychological, moral, but also socio-political sense because the tension is something that needs to be resolved.

My discussion has been rather abstract. I will conclude this section with a brief discussion of an interesting example that illustrates both that people understand their life-world as a narrative and how narrative performance can be used as a source of (new) understanding problems with responsible agency. The example is taken from a research on marital woman battering. For this research Hydén (1994) had an interview with a man who had created his own narrative about his use of alcohol and what it means for his marriage: 'When I come to again, I have often done something violent. Hit Pia (his wife) or trashed the apartment' (1994:58). This elicited the following response from Hydén: 'Then it must be unbelievable luck that Pia is still alive' which she then further explains: 'when you get mad at her and it short-circuits in your head, your body takes on a life of its own and becomes violent. It's lucky that you haven't stuck the bread knife in her, or scissors, or hit her even worse than you have' (ibidem). Interestingly the man immediately denies that this risk would exist: 'I would never hurt her that bad' (ibidem). His narrative tries to make sense of his own responsible agency: as he understands it, whenever he experiences a black-out from the alcohol, his responsibility is suspended. Telling the narrative this way, as Hydén understands him, provides him with an excuse for beating his wife. What is interesting about Hydén's intervention is that in her response, she shows him the implications of this narrative: if he really believes that having a black-out would suspend his responsible agency, this will excuse any act that he would perform during a black-out, even killing his wife. In not accepting this implication, the man shows his commitment to a more complex narrative.

While this example suggests how a narrative performance may help people in addressing and solving problems with their responsible agency, it does not yet show how narrative performances may have the potential to motivate people into addressing and solving problems with *social structures* of responsible agency. I will turn to this issue in the next section.

7.7. Narrative performance as confronting discrimination

With this section, I will conclude my argument that the performance of non-deliberative actions may count as holding people responsible without interfering with their epistemic autonomy. So far, my argument has mostly involved conceptual analysis. In the previous section, however, I expressed the ambition to show what it means, in a less abstract way and closer to actual socio-political contexts, that non-deliberative action may

have an emancipatory function, more specifically: may be capable of unravelling the veil of prejudice and, ultimately, may confront practices of hidden prejudice. In this section, therefore, I will explore a case study on African Americans and how narrative performances might be understood as part of the civil rights movement in the 1960s. But my aim in exploring this case is limited: my aim is *not* to show that non-deliberative action *will* reduce practices of prejudice. Instead, my aim is to *illustrate* how we may think of non-deliberative action in a less abstract way and how it may be argued that they do, or do not, contribute to reducing practices of prejudice. Furthermore, I aim to *make plausible* that such actions require more analysis from social scientists and political philosophers. Before moving on to my case study, I want to comment briefly on my methodological approach.

The struggles of African Americans for their civil rights involve a broad and complex range of literature. Exploring this for my present purpose raises numerous methodological problems that I cannot solve within the limits of my research. I will therefore demarcate my case study in various ways. First, I will analyse it by using the narrative view that was developed in the previous section. Second, I will frame my discussion on how non-deliberative actions might reduce practices of prejudice as a discussion on how social protest might influence public opinion. Third, I will choose photography as the 'focal point' for identifying what might count as non-deliberative action and, as such, might influence public opinion. These choices are open to discussion. Below I will briefly explain my choices.

Applying the narrative view can be done in several ways. Here I will assume that the African American's struggle for civil rights indicates a problem with structures of responsible agency in their societies, at least when it comes to effectively warranting the civil rights of these groups. I will call this the 'inequality problem'. I will also assume that African Americans, at the time of the civil rights struggles of the 1950s and 1960s, were often in social positions that provided them with fewer opportunities to solve the inequality problem than other groups of people. Finally, I will assume that this required African Americans to make appeals, using the most effective techniques available for them to exert influence in the public debate, to those other groups. If we think of this 'appeal' as a narrative, it places African Americans in the role of storytellers and those other groups in the role of 'audience'.

To link 'non-deliberative actions' conceptually to 'reducing practices of prejudice, I take two intermediate steps: first, that non-deliberative actions can take the form of *narrative performances* and, second, that narrative performances in their turn can function as *social protest*. Both steps will help me focus the discussion of my case study: once non-deliberative action is explained in terms of social protest, I can make the next step and explore whether social protesters could influence their audience.

Both steps also have their limitations. For several decades now, narrative theory has entered the social sciences. However, the idea that storytelling might function as

social protest, as what has the potential of changing social structures in society, has received little attention as a domain of research: '[sociologists] have treated stories more as texts to be analyzed for the meanings they express than as social performances that are interactively constructed' (Polletta et al. 2011:110).⁵⁵ Second, while the media coverage of the civil rights movement has often been claimed to greatly impact the 'northern white liberals', detailed and critical analyses of this claim, specifically when it comes to the impact of civil rights photography, seem scarce.

Narrative performance is about telling stories. This will usually be understood literally: *telling* stories, i.e. as verbal acts. Although verbal acts of storytelling already count as non-deliberative, I am interested more specifically in non-verbal acts of storytelling. I will choose photography to explore the possibility of non-verbal storytelling⁵⁶ and more specifically on non-verbal storytelling as social protest because storytelling might perform its emancipatory function as social protest. And I mean to discuss photographs in a broad sense: not just what the photographs show, but also the context in which they function. The photographs are not to be equated simply with the image itself, but they include the actions needed for taking the picture and for getting the picture published in press media. Strictly, a distinction should be made between the non-verbal narrative performances that are being photographed and the performance of photographing itself as another type of non-verbal performance. Although this distinction is relevant, my discussion is to analyse how the combination of both performances may strengthen the narrative implicit in the photographed performance. To put it differently: press photographs may function as the amplifier of that narrative.

Again, this choice for using photography too has its limitations for my case study. First, even researchers who have used narrative theory to study the power of storytelling, for example, in social protest, have mostly focused on 'language-based storytelling, either verbal or written' (Goodnow 2020:266), i.e. on *telling* rather than *showing* stories.⁵⁷ Second, even researchers who do acknowledge the transformative power of social movements and their protests tend to display what McAdam calls an 'ideational bias', i.e. the tendency to 'focus on the speeches, writings, statements, or other formal ideological pronouncements by movement factors' (McAdam 1996a:341), in this way giving priority in their research to 'words' over 'actions'.

The approach, as explained above, is limited to analysing the effects of narrative performance on an audience. This would assume that getting this audience to identify

55 Also see J.E. Davis (2002b); Polletta (2006). While these sources seem to comment mostly on sociological research, the lack of attention for the transformative power of social protest also holds for social psychology. While the perpetuation of injustices and how their victims experience them has been the focus of social psychological research, the 'analysis of resistance has occurred in many forms, largely on the sidelines of mainstream social psychology' (2018a:13).

56 Fisher (1987) allows for the possibility of non-verbal storytelling, but he does not further explore it.

57 This observation has especially been made by those working in visual communication theory: see Goodnow (2020).

with the narrative, with the African American's struggle for civil rights, is enough to motivate them into addressing and solving the inequality problem. In short: to reduce practices of prejudice. This assumption would be too optimistic. A more inclusive analysis than I can offer here should specify what it means that practices of prejudices are reduced – new laws on equality? – and whether a change in the public opinion of a powerful audience has had any effect on this. Again, my aim here is much more modest. I only want to show how we may understand 'non-deliberative action' less abstractly.

One final note on choosing my case study. I had wanted to add a case study on narrative performances in the Roma civil rights movement. But this did not prove to be easy. While much has been written on the civil rights movement as an African American struggle for equality, Roma struggles for equality seem less developed and at least less well documented. Still, in 2010 Greenberg concluded, reporting his findings on segregated Roma education, that '*there is no Roma civil rights movement*' (2010:920, original emphasis). This is too pessimistic.⁵⁸ Even so, the litigation strategy still seems to dominate efforts to realise equal rights for Roma people.⁵⁹ At the same time, Roma people have not yet been part of a civil rights movement that received as much media attention as did the African American civil rights movement. While the case *Brown v. Board of Education* was important, the African American struggle for equal education reached a wider, white audience and became national news when a small group of African American students wanted to claim the rights, approved of by the Supreme Court, and registered for what was still a white school. Among them was Elizabeth Eckford. The Ostrava case has not achieved this same impact. Very different from the case *Brown v. Board of Education*, the *Ostrava* case has to offer only two initials of an otherwise anonymous child, H.D., and no powerful photographic image is available that might show an ongoing struggle for equality of education or illustrate, for my purposes, a narrative performance. Nevertheless, if we look on a more local scale, we do find examples of narrative performances.⁶⁰

Case study: Photographic moments in the African American civil rights movement

My first case study is concerned with the African American struggle for civil rights in the 1950s and 1960s. For this, I will limit my examples to public actions that were captured in photographs that nowadays are thought to be the most powerful and representative images of the civil rights movement. I will specifically focus on the role of photographs in the coverage of the Birmingham campaign. My aim is to explore to what extent civil rights protesters succeeded with their actions in influencing their audience.

The context for the civil rights movement in the United States was a society that is founded on a commitment to the idea 'that all men are created equal, that they are

58 Various studies show how, after the Second World War, Roma people became activists: see Donert (2017); Knesebeck (2011); Beck and Ivasiuc (2018).

59 See Open Society Foundations (2016).

60 See Harper (2012) Orosz (2016).

endowed by their Creator with certain inalienable Rights, that among these are Life, Liberty and the pursuit of Happiness' (quoted in Kluger 1975/ 2004:28), but denied these rights to African Americans, first by recognising slavery as a lawful practice, and later, after the civil war, when slavery was abolished, by creating all kinds of obstacles, both legal and practical. From an early moment, African Americans had challenged, on their own or by organising themselves, the formal segregated and informal racist practices of their society. The narrative that motivated whatever actions they performed might be described like this: we African Americans are prepared to challenge society, even to the point of disobeying its laws, and to demand recognition of our humanity so that we too may enjoy the rights of life, liberty and the pursuit of happiness.⁶¹

In the 1950s and 1960s, various organisations and citizens were involved in the American civil rights movement. To address and solve the equality problem, they explored various strategies. Initially, the most dominant was the litigation strategy, selecting potentially powerful cases and bringing them to court.⁶² It may be considered as a form of influencing that consists of deliberative, specifically highly technical, juridical modes of reasoning, aimed at persuading an audience of judges. Sometimes this strategy was successful. But as Elizabeth Eckford had to experience, juridical victories did not prove to be enough. Even after the Supreme Court had decided the case *Brown v. Board of Education* (1954) in favour of the African American cause, separate public schools remained to exist. African American students now had to enforce what was juridically already their right. It required a change of strategies. As the years went by, protesters began to shift their attention towards strategies that aimed at influencing different audiences, such as public opinion, at both a national and an international level, and politicians and policymakers.⁶³

African American protesters wanted to influence various audiences. For this case study, I will focus on white people from the northern states.⁶⁴ It might be wondered

61 For statements by leaders from the Civil Rights Movements: see, e.g. King (1963). King does not explicitly refer to the rights of life, liberty and the pursuit of happiness or to the Declaration of Independence. I nevertheless think that these rights are key ingredients of this narrative, in whatever way it is formulated. Even as slavery had ended, African Americans still had to fear for their life and liberty: see McGuire (2010). Also, President John F. Kennedy picked up on such rights as he reflected on the position of African Americans in society, for example when he says: 'he cannot enjoy the full and free life which all of us want' (Roberts and Klibanoff 2006:333; also see: <https://teachingamericanhistory.org/library/document/radio-and-television-report-to-the-american-people-on-civil-rights/>, accessed on 31 December 2020). For now, my proposal for the underlying narrative is intended as a guide for my further discussion.

62 See Kolb (2007), Chapter 6.

63 For a discussion of various strategies that were employed within the civil rights movement: see Kolb (2007), Part II.

64 As white people from the northern states were more moderate than those in the southern states, King (1963/ 2000:97) seems to hint at this audience when he remarks that 'the Negro's great stumbling block in the stride toward freedom is not the White Citizens Councillor or the Ku Klux Klanner but the white moderate who is more devoted to order than to justice; who prefers a negative peace which is the absence of tension to a positive peace which is the presence of justice; who constantly says, "I agree with you in the goal you seek, but I can't agree with your methods of direct action"; who paternalistically feels that he can set the timetable for another man's freedom; who lives by the myth of time; and who constantly advises the Negro to wait until a "more convenient season."'

why African American protesters might have wanted to win the sympathy of this audience. Did white northerners not know about the segregation and other obstacles in the southern states that prevented African Americans from being first-class citizens? Or if they did, why they did not do help their fellow citizens? The answer is a complex combination of both. People may sometimes be in social positions, as I suggested earlier, to solve certain problems with structures of responsible agency, but at the same time may lack motivation because they *do not personally suffer* from these problems, or because they *do not even personally observe* them. It also holds for white northerners: they knew something, but they lacked the motivation to explore what they knew, so they remained ignorant. The problem was not that northern white people could not have known about the situation of African Americans in the southern states, but rather what they wanted to learn from it: as Myrdal wrote, ‘The Northerners want to hear as little as possible about the Negroes, both in the South and in the North [...] The result is an astonishing ignorance about the Negro on the part of the white public in the North’ (Myrdal 1944/1962:48). But once Myrdal published in 1944 his *An American Dilemma*, with the subtitle: *The Negro Problem & Modern Democracy*, a very extensive study of race in the United States,⁶⁵ white northerners could no longer deny having any access to what was going on in the southern states. But while it may have helped the white northerners to develop liberal ideas on race issues,⁶⁶ nearly two decades later, they still seemed indifferent: ‘Whites in the North had the luxury of holding more progressive attitudes on race than did whites in the South, because they were born into a society that erected fewer visible barriers to nonwhite advancement and because they had few opportunities to observe the material conditions of black life’ Berger (2011:61). Therefore, it seems plausible that if the Implicit Association Test had already existed, at the beginning of the 1960s, white northerners might have been tested as having an implicit bias.

A broad range of different kinds of actions helped African American protesters to reach their audience. Some of them can be considered at the nexus of rhetorical deliberation and *verbal* narrative performance, such as Martin Luther King’s letter from the Birmingham jail.⁶⁷ But others consisted of what I would consider as *non-verbal* narrative performances: actions whose narrative significance was clear to anyone directly involved in the situation and to the public who was informed of those actions in newspapers and, later, on television. A well-known and prominent example is the action performed by Rosa Parks when she got on the bus and sat down in the section which she knew was perceived as reserved for white people.⁶⁸ The Greensboro sit-ins provide yet

65 It was commissioned to him by the Carnegie Corporation.

66 See Kolb, who observes, in an analysis of changes in American public opinion on equal treatment: ‘A sea change in racial attitudes regarding the principles of equal treatment in a twenty-year period between the early 1940s and the early 1960s’ (2007:156).

67 See, e.g. Fulkerson (1979) and Watson (2004). For more recent references: see Johnson (2007:21–22, note 2).

68 See McGuire (2010). For a discussion of how the act by Rosa Parks may be understood to influence (white) public opinion: Blaakman (2012).

another example.⁶⁹ I would even claim that every protest march by African Americans in those days should be counted as a non-verbal narrative performance. Even while they held up signs voicing their disagreements with segregation, what mattered was the mere fact that they showed up in large numbers and marched through the streets holding up signs: even without reading the signs, it would have been clear to bystanders what the protest march was all about.

As I already indicated, I will focus on non-verbal narrative performances which the audience of northern white liberals could see through the photographic eye. The coverage of the civil rights movement offers many examples. Those that proved to be influential were not necessarily the result of a carefully planned strategy of influencing public opinion. Such was the case with the photograph of Elizabeth Eckford.⁷⁰ It was the result of an unhappy coincidence. It had never been a strategic goal to leave her all by her own, at the mercy of the crowd near the entrance of Little Rock High School. In fact, she and her African American fellow students were supposed to gather at a certain meeting time and place from where, as it has been advised, not their parents but white and African American ministers would accompany them to Little Rock High School. While her fellow students were informed in time, Elizabeth Eckford did not know about this and arrived at the High School all alone.⁷¹ And it had never been a strategic goal that some photographer would capture her in her most vulnerable moment. Even the young woman whom Will Counts' photograph shows as immediately walking behind Elizabeth Eckford, shouting 'Nigger!' at her (Jacoway 2007:6), appeared in the crown only by accident. She was later identified as Hazel Bryan. As Jacoway describes her perspective, she and a girlfriend got 'swept up into the spirit of the moment, joined the heckling', but without 'awareness of the issues involved' (Jacoway 2007:8).⁷² Hazel Bryan 'knew immediately she had made a mistake, and as the Civil Rights Movement unfolded and she understood increasingly the depth of her error, she called Elizabeth and apologized in 1962' (Jacoway 2007:7–8).

Once African American protesters chose to influence public opinion, this was no easy strategy. The audience of northern white liberals was mainly informed by an equally white press. Whatever press photographs were published were taken by white photographers. To understand how African American protesters made a shift in their strategy, moving from litigation towards influencing a variety of white audiences, I will turn to a discussion of the

69 See Polletta (2002) and Abel (2010, Chapter 8).

70 For the impact the photograph and the story of Little Rock High School had in the media: see Roberts and Klibanoff (2006, Chapter 11).

71 See Roberts and Klibanoff (2006, Chapter 11).

72 Yet another example of photographs that impressed audiences were the images of Emmett Till, a 14-year-old African American boy who was tortured and murdered in 1955 after talking to a white woman. Interestingly, in this case, the white audience saw other images than the African American audience: 'While the white press did not publish images of Till's body itself, it did publish searing images of the rituals of mourning that surrounded his death' (Pool 2012:20).

so-called Birmingham campaign.⁷³

The Birmingham campaign was an initiative of one of the national African American civil rights organisations in the United States, the Southern Christian Leadership Conference (SCLC), which was then led by Martin Luther King, its first president. It was set up in 1957 to support local groups of African Americans in their protest actions against segregation. In its first years, it had difficulties in finding the rights ways of effectively supporting local protesters. Its first major opportunity to get involved in a large-scale protest movement was offered by the Albany Movement, a cooperation of the Student Nonviolent Coordinating Committee (SNCC) and the National Association for the Advancement of Colored People (NAACP). The Albany movement had chosen the city of Albany in the state of Georgia as the location for large-scale protest actions against segregation it wanted to support and organise. As the Albany Movement lacked a major strategy, its president, William Anderson, decided to ask for the help of the SCLC.⁷⁴

In the Albany Movement, the SCLC, SNCC and NAACP combined forces to organise in Albany a series of actions for which they employed the strategy of a *community-wide protest campaign* which was ‘not a single new protest technique, but a way of combining previously successful tactics that had become insufficient in themselves to provoke crisis’ (Kolb 2007:118).⁷⁵ Another strategy was that these actions, whatever form they took, had to be non-violent, for principled reasons, but also for yet another reason. By this time, African American organisations had noticed that mainstream media began to take an interest in African American protest actions and that media, also due to the increasing importance of television in the 1950s, were gaining influence in shaping *national* public opinion. And they began to understand the basic logic of media coverage: ‘The more sensational an event, the more likely the national coverage’ (Eskew 1997:23).⁷⁶ Another strategic goal was, therefore, to provoke policemen to take recourse to violence in suppressing and ending protest actions, as this would create a dramatic contrast with the non-violent character of these actions. Protesting for the right to life, liberty and the pursuit of happiness peacefully while being hindered by ‘bad guys’ using violence. What more powerful narrative was needed? If these strategies succeeded, it would provide the bargaining power the Albany Movement needed for overturning segregation policies and practices. The Albany Movement, however, failed to realise its ambition, partly because Albany’s police chief, Laurie Pritchett, was not provoked in letting his policemen use violence but instead instructed them to arrest

73 Kolb (2007) distinguishes between the Disruption Mechanism and the Public Preference Mechanism. While the second strategy assumes that ‘social movements can influence and mobilize public opinion in order to impact public policy’ (2007:140), the first strategy assumes that they ‘can obtain bargaining power by creating institutional disruption through the massive use of direct action tactics’ (2007:115). The Birmingham campaign is illustrative of both strategies.

74 See Eskew (1997).

75 Also, see McAdam (1983).

76 For a discussion of the importance of specifically television news for the civil rights movement: see Bodroghkozy (2012).

protesters who then were taken away to locked up in jails, not those of Albany itself, but of neighbouring counties and cities.⁷⁷

Circumstances in the city of Birmingham in the state of Alabama provided a second opportunity for the SCLCC to organise a campaign. One important difference with circumstances in Albany was that Birmingham had as its ‘Commissioner of Public Safety “Bull” Connor, a notorious racist and hothead who could be depended on *not* to respond nonviolently’ (Hubbard 1968:5, original emphasis). This held the promise of realising their strategic goals and showing to a national white audience their narrative of people who had the courage and determination to challenge violent brutes, denying them what every human being’s right was.

But things were not that simple. The Birmingham campaign was carefully planned to avoid the failure of the Albany campaign: the main goals, strategies and tactics were being revised. Detailed plans were made to prepare for sit-ins. Beforehand lunch counters were selected, even their seats counted, to plan for the most effective sit-ins. Finally, on 3 April 1963, the campaign started with various sit-ins at different cafeterias. Soon the first protesters were arrested. But as the sit-ins continued in the following days, they were losing impact as managers of lunch counters stopped calling the police and instead closed their stores. Meanwhile, the press lost interest in covering the protest events. Plans for alternative protest actions also ran into problems. A temporary injunction raised a legal blockade for street parades and marches. Even when the campaign engaged in new protests, they lacked sensation to attract the attention of the media as “Bull” Connor had resisted provocations to use violence in arresting protesters. What could help the campaign move forward was planning new, somehow more sensational, protests. But for these, the Birmingham campaign needed volunteers. But their human resources began to deplete: ‘We had run out of troops. We had scraped the bottom of the barrel of adults who could go [to jail]’ (quoted in Eskew 1997:261), as Wyatt Tee Walker, a strategist for the campaign, recalled.

The campaign began to falter. But the turning point was what one newspaper would call the Children’s Crusade. Facing what seemed to be a new failure, the Birmingham Campaign shifted to recruiting children: mostly high school students, but even younger children would join the new protests. A march of school children was planned for 2 May 1963. On the morning of that day, thousands of children skipped school and gathered in the Sixteenth Street Baptist Church and other churches. In the early afternoon, children began to leave the churches and walk, in small groups, towards different targets. Towards the end of the afternoon, 959 children had been arrested, peacefully however.⁷⁸ In this way, the Birmingham campaign picked up on a tactic that has been part of the Albany campaign: filling the local jails. The next day, 3 May, the children would again set out to march. But this time, “Bull” Connor was determined to stop the march. He had

77 See Eskew (1997:346, note 53).

78 See Kasher (1996:95). Also, see Tougas (2011).

assembled a troop of policemen, including a squad of six German shepherds, and he had also forced firefighters to be present and bring with them their hoses and tankers: if necessary, they were to spread the crowd. The policemen and firefighters were positioned on the eastern edge of Kelly Ingram Park. As marching children began to show up, a police captain warned them to turn back, but most children kept on walking. Then the firefighters were ordered to use the water hoses. They started to blast water on the children, gradually increasing the water pressure. As Tougas explains: ‘When a fire hose is turned on half strength, the water can knock a person to the ground. At full strength, it can strip the bark off a tree and tear a brick from a building’ (2011:6). It kept on for two hours. As African American bystanders were watching what happened, some of them began to throw bricks and bottles at the firemen. “Bull” Connor ordered his policemen to use their dogs to control the crowd. The demonstration ended around 15.00 o’clock when the leaders of the Birmingham campaign decided to call off the protest actions.

The events of this day were widely reported in national and international media, both in newspapers and on television. In the following days, protest actions continued – successfully, even if this meant getting the jails filled with children – kept receiving media attention. Finally, the Birmingham campaign had succeeded in winning the public opinion’s sympathy.

The extent to which the children’s crusade meant a successful turning point in the Birmingham campaign was due in part, as it is often claimed, to the photographs that were taken and published in the white press. One of them shows a (white) police officer having his right hand in a firm grip on the shirt and sweater of a neatly dressed young African American, while the police dog he is holding on a leash has jumped up and is trying to bite the young man above the waist.⁷⁹ In the foreground, on the right side, another police officer can be seen, also holding a police dog on a leash. In the background, African American bystanders pass along: those that are closer to the scene have turned their heads to see what is happening. The photograph, taken on 3 May 1963 by Bill Hudson, was published ‘the next day on the front pages of dozens of northern papers’ (Berger 2011:36): ‘If there was any single event or moment at which the 1960s generation of ‘new Negroes’ can be said to have turned into a major social force, the appearance of that photograph was it’ (quoted in Berger 2011:64). The young man was afterwards identified as Walter Gadsden.

This was not the only image that impressed the northern white audience. Another photograph was taken by Charles Moore.⁸⁰ It shows three African American people, one young woman and two young men, their clothes soaking wet: all three have turned their back against a strong jet of water coming from the right that presses the young woman

79 The photograph by Bill Hudson is printed both on the front cover of Berger (2011) and as figure 24 (2011:36).

80 For this photograph: see Tougas (2011:32, also printed on the front cover) and Moore (1991:99).

and a tall young man against a facade, a young man in between them attempts to move away and holding his hand to the left of his face to shield it off from the jet of water.⁸¹

Did photographs like these help the northern white liberals to understand the African American narrative and to identify with them? Did they pick up on the narrative performances, as pictured by these photographs, as expressing a problem with America's structure of responsible agency towards warranting equality between white and African American people? And finally, did it motivate them to address and contribute to solving the equality problem? These are hard questions.

The standard view is that the photographs had an enormous impact that helped change public opinion in various audiences: white northerners, but also, to some extent, audiences in the southern states,⁸² politicians in Washington and audiences in many countries around the world.⁸³ The images had succeeded in creating a dramatic tension between African Americans and their white opponents as one between good citizens and bad citizens, between civilisation and barbarism.⁸⁴ The responses of various audiences were emotional: ranging from shock and remorse to shame.⁸⁵ They seemed particularly strong in the case of some white northerners because, as Berger makes clear, '[s]ome of the most forceful and pained to the Birmingham violence came from white liberal intellectuals in the North' (Berger 2011:60). Whatever the impact on other audiences, '[f]or white northerners, the cameras provided incontrovertible evidence of American unfairness, inhumanity, and brutality. Any suggestion by southern whites that their Jim Crow laws and lifestyle were moral, legal, or practical, or that the South's Negroes were fundamentally happy, was demolished by the images' (Roberts and Klibanoff 2006:321). In their turn, these emotions seemed to motivate various audiences into addressing and solving the inequality problem, not the least of them was President John F. Kenney who, on 11 June 1963, who said, in a televised speech: 'The events in Birmingham and elsewhere have so increased the cries for equality that no city or state or legislative body can prudently choose to ignore them' (Eskew 1997:310). These events were the reason for him to announce legislation that would banish racial discrimination: 'I am, therefore, asking the Congress to enact legislation giving all Americans the right to be served in facilities which are open to the public – hotels, restaurants, theaters, retail stores, and similar establishments'.⁸⁶ One year later, this promise was fulfilled when the

81 For a detailed analysis of this photograph and of yet another photograph by Moore, again picturing African American people being the target of firemen: see Johnson (2007).

82 Southern newspapers were reluctant to report about the use of dogs and fire hoses against African American children: see Roberts and Klibanoff (2006); Tougas (2011).

83 See, e.g. Durham (1991); Kasher (1996); Roberts and Klibanoff (2006); Johnson (2007); Tougas (2011).

84 See Johnson (2007). One journalist, for example, Ralph McGill, summarised the actions of policemen and firemen succinctly as 'Birmingham's barbarism' (Roberts and Klibanoff 2006:323).

85 For various examples of such emotional responses: see Roberts and Klibanoff (2006, Chapter 19); Johnson (2007); Berger (2011, Chapter 2).

86 See <https://teachingamericanhistory.org/library/document/radio-and-television-report-to-the-american-people-on-civil-rights/> (accessed on 31 December 2020).

Civil Rights Act was enacted.⁸⁷

President Kennedy already warned against assuming new legislation would solve the problem: 'legislation, I repeat, cannot solve this problem alone. It must be solved in the homes of every American in every community across our country'.⁸⁸ The legislative change may have been the result of the political leverage the civil rights movement created with their large-scale actions, rather than of successfully influencing public opinion and creating majority support: 'the price for the lack of majority support has been the weakening of enforcement mechanisms, which largely determine the real-world impact of legislation' (Kolb 2007:157).⁸⁹

Various commentators seem to halt at observing the emotional responses that photographic images elicited, perhaps because they consider them as sufficient evidence for assuming a change in attitudes or because they consider the introduction of new, anti-discrimination legislation by itself as an important landmark. An interesting exception is Berger (2011), who offers a more detailed analysis of white northerners' epistemic agency. While, for example, Johnson (2007) claims that the Birmingham campaign succeeded in addressing white people's conscience and in shifting the national debate over the inequality problem from *political* to *moral* vocabularies, Berger provides a more nuanced analysis.

Berger explains how white people, particularly white northerners, were not a passive audience in viewing the photographic images of the civil rights events. The photograph of, for example, Walter Gadsden, the young man being attacked by a police dog, was generally interpreted by northern reporters as showing, as Berger summarises it, 'the youth's innocence and composure in the face of unwarranted violence' (2011:36).⁹⁰ He was described with words like 'saint' or 'passive', indicating that white reporters and their audience wanted to see a young man that, calm and resigned (look at his downward gaze!), was walking (look at his left knee!) into the hands of a brute policeman and the jaws of an aggressive police dog. While this interpretation was consistent with the narrative that the Civil Rights Movement, and particularly the SCLC, wanted to convey to white audiences, it was not consistent with what had happened and made the photograph possible. It is right that Birmingham protesters had been trained to remain passive towards brutal acts by the police. But Walter Gadsden, whose family did not sympathise with the Birmingham protest, was only a bystander who had come to look at the protest marches. It was a coincidence that he got caught up in a manoeuvre from the dog squad. While the white northern audience chose its own interpretation, readers

87 See Tougas (2011).

88 See <https://teachingamericanhistory.org/library/document/radio-and-television-report-to-the-american-people-on-civil-rights/> (accessed on 31 December 2020).

89 Kolb (2007) analyses the 'evidence for the influence of public opinion on civil rights legislation' but makes clear that it is difficult to interpret.

90 For examples of how reporters and other (white) people responded to this photograph: see Berger (2011); Roberts and Klibanoff (2006).

of African American newspapers were better informed.⁹¹ They succeeded in identifying and interviewing the young man who commented: ‘When [the policeman] whirled me around, jerking me toward the dog, I automatically threw my knee up in front of the dog’s head’ (quoted in Berger 2011:37) because this was the way his father had taught him to protect him from aggressive dogs. The white audience had also overlooked that small gesture of active resistance: Gadsden’s left hand holding the wrist of the policeman as tight as the policeman holds his sweater.⁹²

Berger also tries to explain the feelings of shame that white felt and expressed in viewing the photographs. In his analysis, it becomes clear that this shame prevented them from being motivated to address and solve the inequality problem. White public responses on the Birmingham events referred in various ways to ‘conscience’ (‘conscience of America’, ‘conscience of the whole people’⁹³). It would indicate that the photographic images succeeded in making white northerners identify with the narrative of African Americans. But it does not explain why they also felt shame. Berger suggests that white northerners still valued their own whiteness (as a moral-political category) or, as I would like to put in my pragmatist vocabulary, that they had not fully integrated their commitments to being white and to ideals of equality. If white northerners did in fact, as Berger suggests, tacitly identify with ‘brutal white policemen’ whose actions they nevertheless condemned, it would explain their feelings of shame which also include a sense of ‘complicity in the crime of segregation’ and therefore of ‘personal failing’ (Berger 2011:63). But that shame prevented white northerners from articulating ‘their own stake in the matter [of racial injustice]’ and from performing ‘the kind of self-examination needed to build empathy and create a coalition of advocates for meaningful social change’ (Berger 2011:75). A self-analysis of their shame, which would mean to uncover they were still committed to their whiteness, was apparently too painful. Therefore, whatever impact their identification with African Americans might have had on their moral motivation, it was attenuated by the impact of their identification with other white people. In short: the voice of their conscience was muffled by their shame.

It is difficult to assess whether the narrative performances of the Birmingham campaign were effective in motivating white northerners to address and solve the inequality problem. My brief discussion of this case study does not support any clear, unambiguous conclusion. But my aim was not to settle this issue. Instead, I wanted to

91 As education and many other institutions were segregated, newspapers too had been segregated. From an early date, African American newspapers had been used to report on segregation and injustices committed against African Americans. See Roberts and Klibanoff (2006). Even when white newspapers also began to report on civil rights struggles, differences remained in how African American and white newspapers reported on them.

92 Critically analysing the civil rights struggle from the perspective of how the press dealt with it, Roberts and Klibanoff (2006) also pay attention to the Gadsden photograph. But even their description fails to observe this detail: ‘[Gadsden] seemed passive, even limp, and stunned to find himself in the clutches of a policeman’s grip and a dog’s jaws’ (2006:318).

93 For these examples, see Berger (resp. 2011:60 and 2011:12).

highlight the possible importance of non-deliberative actions, particularly non-verbal narrative performances, for unravelling the veil of (hidden) prejudice, confronting discrimination and making a case for the need to engage in more detailed analyses of case studies on narrative performances.

7.8. Responsibility for hidden prejudices: criticising the mentalist approach

In chapter 5, I have compared the mentalist and pragmatist frameworks at a *metaphysical* level. For now, I ignore my earlier transcendental argument against mentalism (see section 5.7.) and assume that the mentalist framework may still provide a meaningful view on (responsible) agency. In this section, I will then contrast both frameworks at a *moral-political* level: how do they compare in analysing people's responsibility for hidden prejudices (here, I use 'hidden prejudice' in a sense that is neutral to any framework)? As the mentalist framework dominates debates on responsibility and hidden prejudices, I choose to argue against mentalism and in favour of pragmatism. I will proceed as follows. First, I will argue that the mentalist view on people's responsibility for their hidden prejudice is too narrowly conceived due to features of the mentalist framework itself. Second, I will argue that the mentalist view on holding people responsible for their practices of hidden prejudice is equally narrowly conceived. Third, I will argue that the results of implicit bias tests, such as the Implicit Association Test, whenever they indicate the presence of a hidden prejudice in the tested person, do not force a mentalist interpretation of hidden prejudice *as* implicit bias. This will provide room for a pragmatist interpretation of 'implicit bias' test results.

Mentalism understands the epistemic conditions of *agency* in terms of *mental states*, both conscious and unconscious, and those of *responsible* agency as *conscious, deliberative* and *individual*. The basic mentalist argument for assessing people's responsibility for hidden prejudices is as follows:

1. People may have hidden prejudices (in the sense of 'implicit biases').
2. If they have them, these may cause them to perform discriminatory actions.⁹⁴
3. If they have them, they can have no introspective knowledge of them.

A mentalist analysis of responsibility for hidden prejudice, based on this argument, implies that people are *not* responsible for their implicit biases. It presents the starting view for those who are committed to the mentalist notion of (responsible) agency and the mentalist notion of hidden prejudice ('implicit bias'). Anyone who has these commitments but claims that people *do* have responsibility for their implicit biases is therefore on the defensive: on her rests the burden to show in what way the basic

⁹⁴ They may also influence conscious beliefs. As I am interested only in those implicit biases that have an impact beyond a person's own (conscious) mind, I will ignore this option in my discussion.

argument is incorrect or should be adapted. It might be that the third claim of the basic argument is somehow watered down, for example, by arguing that people might have some control over the effects of their hidden prejudices.⁹⁵ Even then, this will result only in some form of indirect responsibility.⁹⁶ But whatever her suggestions for correcting or adapting the basic argument, her proposal for revising responsible agency for hidden prejudice will remain tied to the contours of the mentalist notions. And this will have the risk that even mentalist reinterpretations of responsibility for hidden prejudices will be too narrowly conceived. To show this, I will first turn to the pragmatist interpretation of hidden prejudice.

Pragmatism understands the epistemic conditions of responsible agency in terms of *rational*, *normative* and *social practices*. I would like to propose the following pragmatist argument, based on my previous pragmatist analyses, for assessing people's responsibility for hidden prejudices:

1. People may perform practices of discrimination.
2. If they do and if they are publicly committed to norms of equality (norms of non-discrimination), a hidden prejudice (in the sense of 'ignorance') may be attributed to them.⁹⁷
3. To the extent that they also perform practices of silencing, their ignorance persists.

Based on this argument, a pragmatist analysis might propose that practices of silencing mark the boundary for responsible agency: people are responsible for their hidden prejudice (ignorance) if, next to practices of discrimination, they also perform practices of silencing.

Perhaps implicit bias may also be viewed as a form of ignorance. But whatever propositional content is hidden in implicit biases and other unconscious mental states, it is not available to the conscious, introspective 'self'. From the perspective of this 'self', a hidden prejudice is the *absence* of knowledge. Instead, in the pragmatist interpretation, ignorance is the *presence* of (self-)knowledge. Ignorance is always tied to knowledge: this is the basic condition for every person. Acquiring knowledge can be a process of mastering and integrating one's concepts/commitments based on some form of a Test-Operate-Test-Exit cycle. Any concept a person has mastered can therefore be described from two perspectives: as what she has already learned and as what she is still ignorant

95 This strategy is pursued by Kelly and Roedder (2008) and Holroyd (2012).

96 See Banaji, Bhaskar and Brownstein (2015), who interpret a variety of philosophical essays (see Faucher 2016; Glasgow 2016; Sie and van Voorst Vader-Bours 2016; Washington and Kelly 2016; Zheng 2016) as expressing this sense of indirect responsibility via a responsibility for remedying the harm: 'though harm due to implicit may be unintended, responsibility for remedying the harm lies firmly with the agent' (2015:184).

97 At this point, their practices may be called practices of hidden prejudice.

about. It means that assessing what people know or do not know also requires an analysis of the context in which people perform their actions.

To analyse hidden prejudice, it is convenient to describe what people know about their practices of discrimination as ignorance. But this ‘ignorance’ is a rather ambiguous term. Ignorance comes in many varieties. To show this, I will suggest various cases of ignorance:

- *The classic liberal*: This person might be someone from the late eighteenth or nineteenth century, familiar with the rights of man vocabulary, and committed to norms of non-discrimination as implying equal freedom (but still in a negative sense), for example, between men and women.
- *The progressive liberal*: This person might be someone from the 1950s; she is familiar with norms of non-discrimination as implying equal rights but understands that this is not enough; she understands that such norms also imply some kind of support, for example, to warrant good education for people that in her society are treated as second-class citizens. Unfortunately, she is not always kind to people of a certain minority and tends to treat them unfavourably in small ways (making jokes about them, interrupting them). In short: she performs what we might call micro-inequities.⁹⁸
- *The willing but inexperienced liberal*: This person resembles the previous person but differs from her in understanding the importance of these micro-inequities. Well, in theory, at least. While she tries hard to change her behaviour, she does not always know *how* to do this.
- *The self-deceiving liberal*: This person is publicly committed to norms of equality (norms of non-discrimination) but, due to her various social positions, lacks the motivation to explore what change this would imply for her.
- *The lonely liberal*: This person has grown up in a community that usually refrains from using discriminative language while talking about certain minorities, as it is familiar with norms of non-discrimination. Nevertheless, this community tends to treat people from these minorities condescendingly. Our protagonist recognises this behaviour but feels uncomfortable with it. She is reluctant to find out about this uncomfortable feeling as she fears it might bring her into trouble and risk the good relations she has with her family and friends.
- *The hypocrite liberal*: This person is publicly committed to norms of equality but is secretly committed to norms of inequality (believing, for example, in the importance of the group she identifies with).

98 See Brennan (2013). For examples, some of which I use here: see Brennan (2016).

All these people can be claimed to be ignorant about their prejudices in some sense, but not the same sense.⁹⁹ What ignorance consists of is different for each case. People may be ignorant of certain aspects of norms of equality (the classic liberal). Or they may be ignorant of the impact of their practices (the progressive liberal). Or even be ignorant about how to perform practices of non-discrimination, and so, while still practising, continues to perform practices of discrimination (the willing but inexperienced liberal).

These various cases indicate that ‘ignorance’ may be used in a descriptive sense – what a person happens to know, and what not –, but also in a normative sense: what people should have known. It might be claimed, for example, that the self-deceiving liberal should know and understand the discriminative effects of her actions. From the perspective of assessing responsibility for practices of discrimination, it may also matter whether people are familiar with a certain norm of equality. To what extent can, in retrospect, the classic liberal be claimed to be responsible for some of her discriminatory behaviour? But it may also matter to which norms of equality people are committed who are in a position of assessing (‘attributing’) prejudice to certain other people, as researchers are when they interpret the test results of their test person.

Each of the cases is formulated *as if* what is morally relevant is only that specific individual and her own actions. But I should stress that, from the perspective of my pragmatist approach, the action of each individual should be understood in the context of social practices. In assessing responsibility for practices of discrimination, the social context may be morally relevant in various senses. One might ask, for example, to what extent the social context in which an individual performs her actions may function as an excuse for holding her responsible (the lonely liberal). But one might also wonder to what extent a certain action can be attributed to one individual. This holds especially for those who work within organisations and may, in some cases, be pressured towards performing discriminatory actions.¹⁰⁰

Having sketched the two basic arguments for assessing responsibility for hidden prejudice, I now turn to a criticism of the *mentalist* basic argument. My aim is to criticise three assumptions of this argument, namely that, in assessing responsibility for hidden prejudices, it suffices that we analyse practices of discrimination as (1) those of an individual, as (2) what is morally blameworthy because of the content of the implicit bias, and as (3) in principle excusable to the extent that they are caused by biases that are unconscious.

The first assumption concerns the methodological individualism that characterises

99 In recent years an interesting field of research has centred on analysing racism and other forms of prejudices in terms of *epistemologies of ignorance*. Contributions in this field show how the ignorance of people belonging to social groups that are relatively privileged (in economic, social, cultural terms) keeps intact and even reinforces existing patterns of societal, discriminatory practices. See Sullivan and Tuana (2007); Gross and McGoey (2015); Kidd et al. (2017).

100 For research that indicates how important an employer’s (discriminatory) preferences can be: see Brief et al. (2000).

research on implicit bias. The mentalist argument pushes the debate on responsibility for hidden prejudices in focussing on individual actions in at least two senses. First, practices with a discriminatory impact are analysed only to the extent that they are reducible to one person. Second, their discriminatory impact is analysed in terms of its epistemic conditions only to the extent that they are reducible to the mental states of that individual. And that is where the analysis stops: ‘even if our attitudes and beliefs come from our culture, they are still in our own minds’, as Harvard researchers on Implicit Bias answer in their FAQ.¹⁰¹ Of course, the reason for these reductions is that the mentalist argument chooses implicit bias as the principal explanation of discrimination. My pragmatist criticism would be that it ignores the social nature of practices of discrimination.¹⁰² For example, discriminatory practices cannot always be reduced to one single individual.¹⁰³ The issue to whom a discriminatory act can be attributed is relevant for a correct assessment of responsibility. Here it is interesting to have another look at the Swedish research on implicit bias of employment recruiters, this time to see how it was set up.¹⁰⁴ The researchers had sent application letters in response to various job ads for two fictional applicants, having the same age, the same work experience, both born in Sweden, but one with a Swedish name, another with a non-native Swedish name (an Arab/Muslim applicant). The researchers then compared how often each of the two fictional applicants was invited for a job interview. The next step was to find for each firm who had been ‘the person who was responsible for selecting whom to invite for interview for exactly the job we applied for in the field experiment’ (2007:6). Finally, these people were asked to take an IAT test. It shows how psychological individualism works in specific research. The procedure to approach firms and select one individual assumes that the results of the application procedure (discriminatory or not) can be ascribed to one individual. It misrecognises, however, that the outcome of an application procedure may also be, and very probably is, the result of the interaction between various individuals within a firm. Matching the result of a firm’s application procedure with only one individual and his or her IAT score, therefore, obfuscates whatever social mechanisms within the firm may have determined the result of the application procedure. This would require, however, a basically other, non-individualist approach.

101 See <https://implicit.harvard.edu/implicit/faqs.html> (accessed on 6 January 2021). The quote is part of their answer to the question: ‘Where do implicit attitudes come from? Is it me or my culture?’ Also, see Leboeuf (2020).

102 For a similar criticism: see Ngo (2016:854), who comments that ‘framing [implicit bias] in terms of the unconscious makes it difficult to give an account of the *uptake* involved in such racist orientations. That is, there is little room for asking how racist stereotypes and attitudes come to be embedded or for clarification of the *role or the participation* of the bias-holder in this process; unconscious or implicit bias confirms that these biases exist but says little about the way they come to be actively embedded in our ways of being’.

103 See Marvasti and McKinney (2007).

104 Rooth (2007): it was already mentioned in 1.2. and discussed in 6.4.

The mental argument also ignores the issue of *how* practices of hidden prejudice, and therefore of how hidden prejudices themselves, persist. This is an issue that can be only analysed by treating practices of discrimination as a *social* phenomenon. In chapter 4, I have suggested that the notion of ‘practices of silencing’, analytically distinct from practices of discrimination, might help to gain more insight into issues of the social ‘reproduction’ of practices of discrimination.¹⁰⁵ A closer analysis of practices of silencing might also reveal that the motives people have to persist in their practices of discrimination may be more complex than a straightforward ‘negative association’ regarding people from minorities.¹⁰⁶

Next, the mentalist argument ignores how the social dimension of practices of both discrimination and silencing already operates from within people’s minds (and of course, I mean this in a pragmatist, not a mentalist, sense!) in a more complex way than the notion of ‘implicit bias’ is able to reflect. In this notion, the social dimension is captured only in terms of an associative or propositional link between a certain group and certain negative characteristics. But the content of an implicit bias does not seem to refer, in any way, to the person having the bias: researchers and theorists on implicit bias do not assume that the content of an implicit bias in terms of ‘I have a negative association...’. It means that when, for example, they discuss implicit biases against African Americans and analyse cases of both white people and African Americans having such implicit biases, they will assume that the content of this ‘implicit bias’ is the same. My pragmatist approach would suggest that we understand the social dimension of a hidden prejudice in terms of implicit, social intentionality: as we-commitments (see sections 6.4. and 6.5). It would allow an understanding of prejudices as both hidden (as ignorance) and as referring to themselves: it

105 To my knowledge, the research on implicit bias is missing any notion that resembles ‘practices of silencing’ and any explanation of the persistence of hidden prejudices.

106 Some social psychologists have explored how people might acquire their implicit biases. For brief discussions of research: see Olson and Fazio (2002); Rudman (2004a); Rudman (2004b). Also see: Nosek and Hansen (2008). One of the suggestions has been that people’s implicit bias might be influenced by ‘cultural biases’. But this research usually focuses on statistical relations between groups (for example, their status) and their tested implicit attitudes (for example, implicit bias). Such research suggests that ‘cultural biases inform implicit attitudes more than explicit attitudes’. The problem with such research is that it ignores the issue of *which* practices, and how, may have contributed to the conveyance of prejudices. But social psychological researchers will not likely look for such practices as their mentalist framework pushes them towards understanding the acquisition of cultural biases in terms of cultural conditioning. For a description of such research: see Olson and Fazio (2002). Such conditioning experiments, like the Implicit Association Test, consist of showing the test subjects a series of words and images. It reveals, on a more subtle level, that social psychologists are committed to the mentalist framework and its ocular metaphors, singling out ‘seeing’ as the most important way to how we ‘acquire knowledge’ (if ‘being conditioned’ deserves this name). Also, see my chapter 1, note 51. (For a brief discussion on this view of ‘perception as continuous with cognition’: see Durrheim 2012.) Interestingly, while I could find research on implicit biases against blind people, I could find no research on blind people having implicit biases. In research on how blind people acquire their racist prejudices, Obasogie (2010) concludes, in my view going against the grain of research on implicit bias, that ‘[t]he significance that society attributes to visual aspects of race comes less from any obvious or self-evident physical differences and more from how social practices train individuals to look differently on certain bodies’ (2010:597).

would imply that, even when African Americans themselves have certain hidden prejudices against African Americans, these hidden prejudices would have a different structure. Once we understand hidden prejudices as part of people's identities, we may better understand how they may fulfil an *epistemic* function in preserving a certain social structure.

These pragmatist criticisms join wider criticisms against methodological individualism in social psychology. For example, Henriques (1984) has criticised social psychological research on racism because it remained committed, in its theoretical frames, to an individual-society dualism.¹⁰⁷ The basic strategy of social psychology has been, he explains, to start from psychological individualism and to treat prejudice as only a problem with the 'rational mechanisms of the mind'.¹⁰⁸ It allowed social psychologists, on the one hand, to ignore the issue of the socio-historical (re)production of racist prejudices, on the other hand, to make a sharp boundary between process (the 'mechanisms' of the mind, those that underlie for example perception and memory) and content (prejudice). The consequence, however, was that practices of discrimination were theorised as only the result of individuals in terms of 'individual aberration and 'irrational behaviour' while society itself was 'assumed to be basically unproblematic' (1984/ 1989:60). This denies the complex relations between an individual having prejudices and her social context.¹⁰⁹

Proponents of the notion of 'implicit bias' have not been deaf to criticisms of ignoring the social dimensions of practices of hidden prejudice.¹¹⁰ For example, Brownstein responds to criticisms that research on implicit bias ignores the non-psychological, i.e. institutional-structural features of society as possible causes of practices of discrimination. Although he welcomes structural criticisms as an invitation to explore the interaction between social structures and psychological attitudes, he argues that proponents of a social-structural approach have nevertheless difficulties with explaining cases of inequality in which psychological attitudes (for example, high or low levels of racial bias) have an explanatory role, independent from structural features of those cases.¹¹¹ But this response

107 Although his analysis dates to the early 1980s, when the research on implicit bias still had to start, I think the analysis still holds. See this section, note 98. For a more recent criticism: see Adams et al. (2008b).

108 While Henriques claims that this basic strategy underlies social psychological research at least as early as the 1950s, Stroebe and Insko (1989) suggest that 'the watershed came with Tajfel's (1969) paper, "Cognitive Aspects of Prejudice," in which he formulated what were to become the basic postulates of the "cognitive approach" to stereotypes and prejudice. The role of motivational biases was minimized, and stereotypes were conceived of as categories that bring coherence and order to our social environment. The biases and exaggerations characteristic of stereotypes were seen as the result of limitations of the human capacity for processing information' (1989:5). For critical comments like Henriques's criticism of the individual-society dualism: see Bader (1995:16–20).

109 To my knowledge, this criticism is not discussed within the research literature on implicit bias. More recently, Leach (2002) has observed that Henriques's criticism still holds, also for the newer research on implicit bias. Also, see Salter and Adams (2013).

110 See, for example, Jost (2018), although he does not address the specific criticisms that I mentioned against methodological individualism.

111 Brownstein (2020) is specifically referring to the findings of research by Chetty and colleagues. For a recent publication: see Chetty et al. (2020).

ignores other social dimensions of practices of discrimination. Contrasting individual and social approaches is not just about contrasting psychological and structural explanations, but also, for example, of finding out why and how certain attitudes persist (see my earlier remarks). It also ignores that social circumstances ('outside the head') are a source not only of *explaining* (namely, where prejudices in one's head come from) but also of *holding responsible* or *excusing*.

The second assumption concerns the basically empirical approach of the mentalist argument. It is problematic on a meta-ethical level. Researchers and ethicists generally agree that the *content* of an implicit bias by itself represents what is ethically problematic and that if the implicit bias causes a person to perform a discriminatory action, this action too should count as ethically problematic. Consequently, the assessment of responsibility becomes an issue of whether she has control over her hidden prejudices (reflective volitional autonomy) or whether she (consciously!) knows about her hidden prejudices (reflective epistemic autonomy). Plausible as this reasoning seems, it risks reducing assessments of what is morally problematic and of responsibility to factual assessments, namely of uncovering whether any implicit bias was at work. This would ignore possible other causes (social pressure, for example, as in the case of the lonely liberal). But this would also, to borrow an argument from Smiley, 'not only conflate causation and blameworthiness into a single moral term but treat the two as part of one moral fact about an individual's moral agency' (1992:75). While Smiley makes this point in a different, broader context – in an argument against the modern (in my view: mentalist) notion of moral responsibility –, her criticism also applies here, namely that we cannot disconnect (moral) norms from the community that is committed to these norms.¹¹² I do not mean to claim that (moral) norms are 'relative', but simply to draw attention to at least two implications of this idea. First, it implies that the content of an implicit bias cannot be described independently from the norms of those who describe that content. Second, it implies that a person may perform practices of discrimination, not necessarily because she is prejudiced (although she may have been tested for having an implicit bias), but because the community to which she belongs does not agree on norms of non-discrimination¹¹³ or because people are familiar with norms of non-discrimination but fail to understand what it implies for their own actions.¹¹⁴

112 I want to remind of MacIntyre's diagnosis that '[c]ontemporary analytic philosophers [...] often take themselves to be representing the timeless form of practical reasoning as such, when they are in fact representing the form of practical reasoning specific to their own liberal individualist culture' (1988:340). I think this may also hold for social psychologists.

113 See Eidelson (2013).

114 The pragmatist view on the use of norms of non-discrimination (even their failed use) is, of course, one of application *and* alteration (see my analysis in section 7.4.). It would be interesting to know, for example, how researchers on implicit bias understand applications of norms of non-discrimination (and other concepts): as application-and-alteration or as a 'two-phase story'? This is morally relevant because the application-and-alteration is one allows to think of (moral) norms as involving a learning process. This is not clear in analyses of implicit bias: does it allow for the possibility (as in the case of the willing but inexperienced liberal) that a discriminative action may be the result of a learning process rather than of a hidden prejudice.

My pragmatist account proposes to reject an understanding of the blameworthiness of discriminatory practices in terms of an underlying implicit bias, and instead to favour a broad, contextual understanding of these practices, which takes into account, for example, norms about what people should know and possible disagreements on (moral) norms.

The third assumption concerns the view on ‘hidden’: this is important both for explaining what makes prejudices ‘hidden’ and for assessing responsibility for hidden prejudices. As with the other assumptions, the mentalist and pragmatist approaches have fundamentally different views. While the mentalist view denies that people have introspective access to their hidden (‘unconscious’) prejudice,¹¹⁵ the pragmatist view claims that people have inferential self-knowledge. Such knowledge is mediated by the actions a person and people from her community perform in various contexts. These actions help people to gain insight into the meanings and the norms of the community’s actions, also those of discrimination or non-discrimination. The mentalist and pragmatist approach may, in some cases, agree that prejudices are hidden from people themselves and that people are not responsible for them. But they will disagree on the reasons for not being responsible. For the pragmatist view, the attribution of ‘hidden prejudice’ (ignorance) is a convenient tool for what might be very different cases, both in how the ignorance should be described and in how responsibility for this ignorance should be assessed. Earlier I provided various cases that represent a selection of these shades of ignorance. It allows the attribution of responsibility in varying degrees. The mentalist view, however, cannot account for what might be morally relevant differences between these cases. Instead, the mentalist view offers only two possibilities: having an implicit bias or not. And therefore, the liberals from my cases, depending on how they would test on implicit bias, would count as implicitly biased or not.¹¹⁶ The mentalist view is therefore too narrow and excludes too many cases of possible responsibility.¹¹⁷

The conceptualisation of ‘hidden’ as ‘unconscious’ is problematic in yet another sense, especially if we look at the research on implicit bias in its American context. The technique of implicit measures, as in the Implicit Association Test, is generally assumed to bypass the problem of socially desirable answers provided by test subjects. But it also has led researchers on implicit bias to assume that the responses, which test subjects cannot control, warrant a mentalist interpretation and therefore indicate the presence of unconscious mental states in the test subjects’ heads. Considering that much of the work on implicit bias has been performed by social psychologists in the U.S.A. and has included

115 As Holroyd (2012:292) explains, this is the whole point of implicit biases: ‘Were we able to detect implicit biases by means of introspection, we would not need such sophisticated indirect measures to discern the presence of such biases’.

116 For a similar criticism of this binary logic: see Kahn (2018). Of course, the underlying problem is the mentalist division of mental states into either conscious or unconscious. I will get back to this in chapter 8.

117 For a similar criticism: see Selmi (2018), although he is referring specifically to legal responsibility.

research on implicit bias on race, this interpretation is surprising.¹¹⁸ Earlier I referred to observations that, during the second half of the twentieth century, prejudices changed in the U.S.A. (and in other countries), transforming from old into new prejudice, from overt into covert prejudice. Was this also a change from ‘conscious’ into ‘unconscious’ prejudice? Might it be the case that prejudiced people have learned to adapt to a society that condemns certain talk about other races?¹¹⁹ If that were true, then instead of treating people as non-responsible for their implicit bias, we should treat them as bigots.

The mentalist argument can also be criticised for the view it implies on *holding responsible*. As we have seen, the starting view is that people are not responsible for their hidden prejudice. It raises the question of what intervention is morally justified to reduce practices of hidden prejudice. May prejudiced people be forced to swallow an anti-prejudice pill? Answering a question like this depends on our assessment of the relation between people and their practices of discrimination. We have already concluded that they are not responsible for these actions. Therefore, whatever intervention we choose to reduce their practices at least does not count as *holding them responsible*. But how then should we understand the relation between these people and their actions and, consequently, what intervention would it justify? We may compare this relation to the relation between people having Tourette’s syndrome and, in some cases, their cursing behaviour.¹²⁰ Even if we think that cursing is inappropriate for moral or other reasons, we do not hold them responsible for it because we think, in their case, it is an involuntary action.¹²¹ While we cannot hold people responsible for involuntary actions, we might nevertheless be justified to intervene if the impact of their involuntary actions interferes with other people’s wellbeing. This might be the case if, for example, a person turns into an involuntary murderer while sleepwalking. Murder, or the risk of murder, would morally justify an intervention.¹²² If the case of the sleepwalking murderer is analogous to the case of an unconsciously discriminating person, then practices of hidden prejudice would also morally justify an intervention to prevent that person from further discrimination. This would still leave many other questions. Would any intervention be justified? Would force be justified?

118 For a similar observation: see Banks and Ford (2009:1058), who comment that ‘the IAT could defensibly be viewed as a subtle measure of conscious psychological processes, of attitudes and beliefs that are known to oneself yet intentionally concealed from researchers. This empirical ambiguity has been practically eclipsed by the unconscious bias account. Why?’

119 See Blum (2002); Billig (2012).

120 See Kushner (1999).

121 See Buss (2012), who comments on their responsibility (here: autonomy) for their involuntary behaviour: ‘People with Tourette’s syndrome, for example, compare their motor and vocal tics to a sneeze. Even though they have some control over when they stop resisting the urge to move or shout, this urge eventually overpowers their resistance. It is precisely for this reason that the Tourette’s tics shed no light on what is special about autonomous action: they are not under the agent’s control because they are not even voluntary; they are not autonomous actions because they are not actions at all’ (2012:681, note 76).

122 This is no thought experiment. For cases of violence during sleepwalking and a nuanced analysis of legal responsibility for such acts of violence: see Popat and Winslade (2015).

The basic argument allows for yet another option: if we adjust it in certain ways, we might conclude that people are, under certain circumstances, responsible for their practices of discrimination. In these cases, an intervention might count as holding responsible. But not any intervention would be justified, for example, an intervention that would violate people's epistemic autonomy. If administering an anti-prejudice pill would be justified, at least it would not be justified to manipulate them into taking this pill.

Whatever any mentalist account would claim on people's responsibility for hidden prejudices or on what intervention would be morally justified, its mentalist assumptions would force them to focus on those kinds of interventions that target the implicit bias by either weakening, disrupting or controlling it.¹²³ And it would force the design of the intervention to focus on influencing *intrapersonal*, i.e. biological, neurological mechanisms of a person's body or brain.¹²⁴ The reason for this is that the mentalist argument leads to an epistemic stalemate. While acquiring knowledge has traditionally been viewed as important for moral development, this road has been cut off because, from the mentalist perspective, acquiring knowledge, to the extent that it is deemed relevant to being responsible,¹²⁵ is identified with deliberative rationality, while the impact of implicit biases on people's conscious mental states is not understood in terms of deliberative rationality. It means that acquiring (self-)knowledge has no emancipating force. One tendency, therefore, is to design interventions – what Brownstein (2017) calls 'change-based interventions' – that 'condition' people in losing negative associations.

In recent years such interventions have been widely criticised for 'medicalising' the approach towards discrimination.¹²⁶ But apart from such criticisms, my pragmatist approach calls for analysing practices of prejudices as *social* practices and for identifying yet other types of practices, such as practices of silencing, that might explain the *persistence* of practices of prejudices. If this social approach is correct, then we may suspect that individual, medicalised interventions will fail to work as long as practices of silencing (or practices with a similar function) continue to exist. Even more so, I suspect that administering an anti-prejudice pill, during the second half of the 1990s, to the

123 For examples: see Brownstein (2017, Section 4.2).

124 As a result, social psychology is marked by a tendency to view racist (and more broadly, prejudiced) people as forms of disease or as pathological cases. Exemplary for such views is the title of the TedxTalk, given by Sylvia Terbeck (earlier mentioned as one of the 'inventors' of the anti-prejudice pill, in chapter 1, section 1): *Prejudice: Can We Cure It?*

125 In the mentalist view, it is possible that people acquire knowledge in a non-deliberative sense, but this would have no relevance for the assessment of their responsibility.

126 For criticisms: see Wellman (2000); Thomas and Brunsma (2014); Kahn (2017) and (2018). Also, see Gilman and Thomas (2016), who sustain their criticism with an extensive reconstruction of the history of how racism began, from the nineteenth century onward, to be pathologised. Scheffler (1992) suggests a relation between this pathologisation/ medicalisation and political liberalism. Interestingly, also outside the psychological research on racism, psychologists, philosophers, and others have been tempted to use notions from pathology and apply them to 'normal' people, as we can see for example in the literature on introspective self-knowledge where 'confabulation' is used to explain the failure of introspective self-knowledge in non-pathological cases (also see chapter 4, note 77).

heads of schools and the psychologists of Ostrava would not have prevented the Ostrava case (see chapter 1, section 1). The anti-prejudice pill is effective because it may reduce amygdala activity in the brain and therefore reduce the negative emotions that a person experiences in looking at faces from people who are the target of her hidden prejudices (see chapter 1, section 2). The problem, however, is that what caused the discrimination of the 18 (and probably many other) Roma children of Ostrava were *dispositions of prejudice* that had solidified – not merely, perhaps not even mainly – in habits of emotional association, but also in inferential habits of thinking about Roma – habits of making uncritical assumptions about their culture, their living conditions, their abilities –, furthermore in social, habitual practices of silencing and, finally, in the externalised (institutionalised, formalised) ‘products’ of such habits: rules, procedures, protocols, psychological tests that – viewed by themselves, i.e. in abstraction from the context in which they are in fact used – are neutral, i.e. do not bear any mark of discrimination. But, once they are applied in certain contexts, they turn out to have a discriminating effect.¹²⁷ Therefore, when prejudiced people experience emotions that have become less negative, for example, in taking a psychological test or deciding on placement for a special school, it will not alter the discriminative effect because inferential habits and the externalised products of prejudiced habits will have remained intact. What it takes to confront such patterns of prejudice is that prejudiced people are *motivated* to change their inferential habits, being open to criticism, and change their rules, psychological tests, et cetera. For this challenge, mentalist interventions, such as the administration of an anti-prejudice pill, fall short. They may reduce negative emotions, but they cannot *motivate* prejudiced people into unravelling their veil of prejudice.

Or can they? Of course, we can think of thought experiments in which anti-prejudice pills do motivate prejudiced people into critically exploring their habits of prejudice and their products. Only empirical research can reveal whether, someday, pills can have that effect. But if it would, again, we would stumble upon the moral problem of manipulation. Would it morally be justified to manipulate people, to secretly administer this new kind of anti-prejudice pill, if this would reduce their practices of prejudices? If we think it is not morally justified, we might instead *recommend* people, *persuade* them, into taking this pill. But what if persuasion does not work?

Drawing on fundamental differences between the mentalist and pragmatist approaches, I have now offered various reasons for rejecting the mentalist account of responsibility for hidden prejudice. Very briefly, its problem is that it understands both hidden prejudices and responsibility too narrowly. For this reason, it is at risk of ignoring what might be relevant moral dimensions or aspects of specific cases of discriminatory practices.

It might nevertheless be claimed that, despite these criticisms of the mentalist

127 I understand such ‘institutionalised products of emotional and inferential habits’ in a broad sense. Such a product may even be the location of people’s homes within a certain region or city – people from different social groups living in different neighbourhoods. See, e.g. Rothstein (2017).

account, the concept itself of implicit bias still holds. Whatever interpretation my pragmatist approach has to offer for hidden prejudices, it is already well behind as the mentalist interpretation is backed up by a growing body of empirical evidence. But that claim would be premature. We should not confuse what functions as an explanation and what needs explanation. I want to make a sharp distinction between *having an implicit bias*, in the mentalist sense of having an unconscious prejudice, and *having an implicit bias test result* (abbreviated as: IB test result), as having one's test indicating a hidden prejudice. In the discussion on implicit bias, both popular and academic, having an implicit and having an IB test result are usually conflated and treated as interchangeable phrases. My claim will be that having an IB test result does not necessarily imply that people have an implicit bias: an IB test result *does not force* a mentalist interpretation of hidden prejudice. I will discuss two different cases.

The first case is that a person has participated in a test for implicit test, for example, the Implicit Association Test, and that her test indicates an IB test result. The attribution of an implicit bias based on an IB test result has mostly been driven by theoretical concerns intrinsic to the mentalist framework to which social psychologists are committed. And although the theoretical construct of an 'implicit bias' has helped them to explain their test results, the IB test result does not, by itself, force a mentalist interpretation. I want to home in for a moment what it means that, according to the research on implicit bias, someone 'has an implicit bias'. If a person has a mental state that is called 'implicit bias' (as indicated, for example, by an IAT-test), this does not mean that this person, for that reason, has a hidden prejudice. Strictly, what an IAT-test (and other similar tests) is testing, is a person's behaviour: the time it takes her to respond to associating various words and pictures with each other. If a person is faster in associating Good/European Americans than she is in associating Good/African Americans, this is supposed to indicate a mental state that is called 'implicit bias' and which exists, consistent with the mentalist view, within the person's head. So, this is *the mentalist notion of implicit bias*. Of course, this mental state is not really an implicit bias. The research on implicit bias attempts to predict discriminatory behaviour. Unless it is claimed that these predictions are faultless, it implies that sometimes people will have a mental state that is called 'implicit bias', and *not* perform discriminatory actions, while sometimes people will *not* have a mental state that is called 'implicit bias', and still perform discriminatory behaviour. What decides whether the IAT-test is good at predicting discriminatory behaviour? For this, yet another experiment is needed that somehow shows whether people perform, or do not perform, discriminatory behaviour. Comparing the results in a statistically relevant sense will then show the statistical value of IAT scores. Having a certain IAT score does not imply that one has (or has not) an implicit bias. It only implies that one has a mental state that represents a statistical tendency to perform a discriminatory action. This is, therefore, *the statistical notion of implicit bias*. It does not require that we understand the mental state that is called

‘implicit bias’ as really representing an implicit bias.

What research on implicit bias shows is that we are unconscious of our reaction time in making certain associations. What it does *not* show is that hidden prejudices are unconscious. That implicit biases are unconscious is part of the definition of the explanans: the notion of implicit bias. And second, if my earlier remarks are correct, it is not even clear how we may attribute to a specific individual an implicit bias because ascribing a mental state that is called ‘implicit bias’ is not yet ascribing an implicit bias. Once we shift from a mentalist towards a statistical notion of implicit bias, we might still claim that the implicit bias that certain people are likely to have will be unconscious. But for what reason? It is sometimes claimed that the IAT-test is resistant to faking. But if implicit biases really are unconscious, it would not be necessary to construct a test that has this insensitivity. Why do researchers on implicit bias not claim that their research is able to ‘unmask’ the hypocrisy of certain prejudiced people?¹²⁸

The problems with possibly conflating the mentalist and statistical notions of prejudice and of interpreting prejudice as either conscious or unconscious indicate, in my view, a problem with the theoretical position of social psychology itself. Bourdieu would analyse its position as one of taking a *stance of objectivism* towards its subjects (chapter 6, section 3). It means that social psychologists must meet the challenge of how to move from the statistical regularities in their observations to an explanation of those regularities. According to Bourdieu’s analysis, they are at risk of sliding from the model of reality to the reality of the model: of moving from measuring ‘reaction time responses’, as indicating an implicit bias, towards taking the mental states which are called ‘implicit bias’, as what really represents an ‘implicit bias’.

A second case is a person, again with an IB test result, but also with a test result from a different kind of experiment that does not test on associations but on actual behaviour. I have already referred several times to the Swedish experiment that combined both tests (Rooth 2007). But would the combined test results of such experiments finally warrant the attribution of an implicit bias? Yes, it would allow the mentalist interpretation of a hidden prejudice, but, as I hope to have shown so far, it would not force it.¹²⁹ And this is all the room I need for claiming that an IB test result might also allow for a pragmatist interpretation of a hidden prejudice.

128 ‘Do Implicit Measures Provide A Window to the Unconscious?’ In answering this question, Gawronski (2009) acknowledges that the methodologies for assessing implicit bias do not unequivocally confirm that they are unconscious.

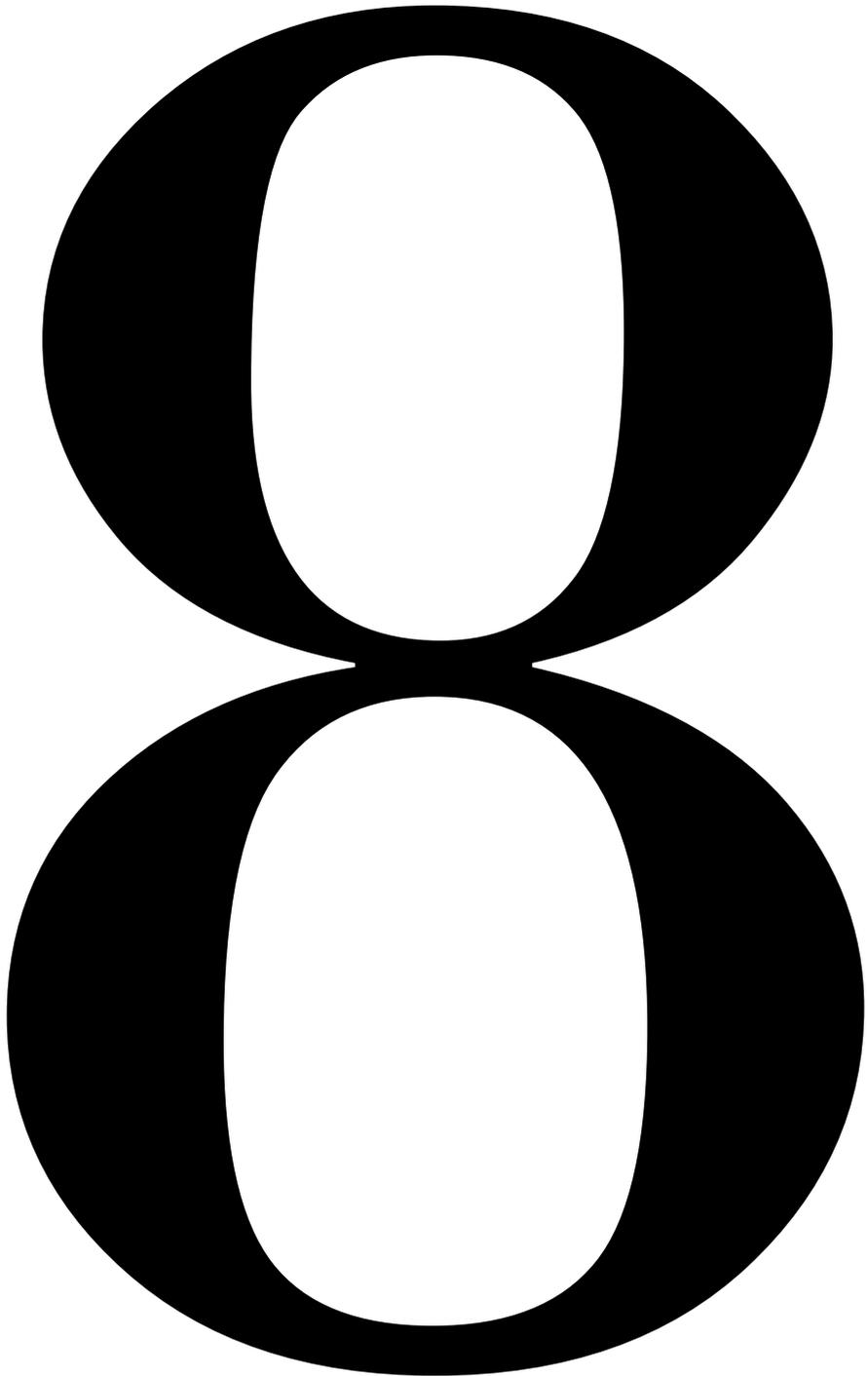
129 For a similar criticism that IB test results do not force a mentalist interpretation: see Selmi (2018).

7.9. Conclusion

I have shown that the Brandomian-pragmatist framework is able to sustain alternative notions of self-knowledge, freedom and responsibility. First, self-knowledge is inferential rather than introspective: people have self-knowledge to the extent that they grasp the implications of their actions, that they can understand these implications from other perspectives and that their understanding is implicit or explicit. It follows that inferential self-knowledge is not the opposite of ignorance. On the contrary: ignorance is always already intrinsic to having inferential self-knowledge. For the liberal-mentalist notion of responsibility, authorising presupposes self-knowledge of one's 'real' self. For the Brandomian-pragmatic notion, this order is reversed. Acquiring self-knowledge (again: in a broad, social sense) is first a matter of authorising. One implication is that uncovering one's self cannot precede one's position in society but is already a product of, and a response to, that position. What helps the Brandomian-pragmatist framework in avoiding the pitfall of social determinism is to take deauthorising as central to the notion of freedom at the moral-political level. It enables people to adopt a critical stance towards the socio-political order. The two notions of inferential self-knowledge and deauthorising can next be used, as I argued, to sustain the claim that people are responsible, in various degrees, for their hidden prejudices. Although this responsibility cannot be assessed without considering specific social contexts, at least it may be claimed that in some contexts, people are responsible for their hidden prejudices.

As the Brandomian-pragmatist notion of non-deliberative action was still too abstract and too far removed from the socio-political contexts of practices of hidden prejudice, I suggested a strategy for adapting this notion to socio-political contexts: it will consist in understanding *inferential rationality*, as what underpins non-deliberative actions, more specifically as *narrative rationality* and therefore to understand non-deliberative action as narrative performance. To show how this strategy might work, I discussed a case study on the news photographs in the civil rights movement.

With this chapter, my aim has been fulfilled to present an alternative, non-mentalist framework that sustains a different, what might be called a liberal-pragmatist, notion of responsibility. This notion promises to provide a different answer to my research question because it draws a different, more ambiguous boundary between self-knowledge and agential self-ignorance and because it implies a somewhat different view on authorising at the socio-political level. While people may be claimed to be responsible for their implicit biases, depending on the context, this does not yet answer my research question. In the next and last chapter, I will turn back to the research question and evaluate what the liberal-mentalist and the liberal-pragmatist notions of responsibility imply for it.



Chapter 8

Conclusion

8.1. One research question...

The photograph of Elizabeth Eckford should remind us that being a citizen in our western, pluralist democracies is no guarantee for enjoying equal treatment. Not in the past. And not even now. Sadly, it seems that in recent decades the desegregation for which African Americans struggled is being reversed by developments of resegregation.¹ In Europe, specifically in the European Union, various audiences do not even seem to be aware of any struggle for equality by the Roma people.² What should we do to reduce practices of prejudice?

The municipality of Amsterdam has expressed ambitions to reduce practices of hidden prejudice. In 2019 it announced a strategy to promote awareness of diversity and inclusion among its personnel, especially its higher management. As part of this strategy, the higher levels of management would receive a training, including an Implicit Association Test: ‘The training provides insight into one’s own prejudices and contributes to awareness and opportunities to recruit without (unconscious) prejudices’ (Amsterdam 2019:10). What will be the next step?³ A policy for letting them swallow an anti-prejudice pill? Moreover, what if managers refuse to take a pill, perhaps for health reasons, or perhaps simply because they disagree? Will they be obligated to – or worse: be manipulated into – taking an anti-prejudice pill? However, what would be the alternative? Situations and questions like these motivated me to explore the following research question:

What are the conditions for non-deliberative actions to count as holding responsible those people who perform practices of hidden prejudice?

My attempt to answer this question led me into domains that are far removed from the struggles of those who experience practices of prejudice: manipulation, determinism, rationality, et cetera. I think it was nevertheless a necessary detour because answering my research question depends, as I suggested in the first chapter, on what framework one chooses for understanding responsible agency. In the previous chapters, I have addressed and contrasted two different frameworks for understanding responsible agency: a mentalist and a pragmatist framework. Each of them carries different assumptions on what rationality means and what it implies for agency and responsibility. My discussion

-
- 1 See e.g. Orfield (2001); Pettigrew (2008). For current segregation in other domains of society: Alexander (2010/ 2012); Quillian et al. (2017).
 - 2 Let me put it this way: the interest in research into implicit bias against Roma people has not been overwhelming.
 - 3 As the strategy of the municipality of Amsterdam makes clear, the next step would be a consciousness-raising training (Amsterdam 2019:10). However, this step is inconsistent with the findings of research on implicit bias. See Woods (2017:640): ‘if racism is so deeply ingrained as to constitute the unconscious, then why would we expect a program of rational consciousness-raising about implicit bias to effectuate changes in the unconscious?’

has been critical of the mentalist framework. It functions as the dominant theoretical background for epistemological and psychological assumptions in debates on people's responsibility (for their implicit biases and, more generally, for their desires, their reasons and their actions) but also, largely implicitly (or should I say: unconsciously?), in liberal-political debates on (in)justice, discrimination and exclusion. My discussion has been constructive in explaining the pragmatist framework because it is lesser known and still underdeveloped. I also assumed that the pragmatist framework is able, whereas the mentalist framework is not, to make conceptual room for understanding non-deliberative, influencing actions as holding people responsible for their practices of hidden prejudice. It is time now to evaluate whether I succeeded in realising my research aims. I will first briefly answer the research question from the two perspectives I discussed: the mentalist and pragmatist perspectives. As these perspectives are based on two frameworks, I will next compare and evaluate both frameworks. Finally, I will make suggestions for future research.

8.2 ...and two answers

Exploring two frameworks has resulted in two different answers to the research question. Before summarising them, I will briefly explicate the basic reasoning that guided my search for answers.

The reasoning is as follows. People are moral-political beings. A notion of *responsible agency* should help us to understand what the conditions for actions are to count as ones for which people are responsible (in a moral or political sense). Such a notion could then function, first, as a norm for assessing in what cases people are responsible for their actions and therefore may be *held responsible*, and second, as a norm for assessing what forms of holding responsible are justified and which are not. The first function seems to be disturbed by hidden prejudices in a sense explained to us by social psychologists, i.e. by influencing certain processes in the brain: if people act under the influence of hidden prejudice, they may not be responsible for their actions. Administering an anti-prejudice pill might solve the problem of hidden prejudices. But this type of action seems to be disturbing the second function. As such, a pill would influence the person in the same way as hidden prejudice does, i.e. by influencing certain processes in the brain, should we not describe such an action as manipulation. However, describing this action as manipulation is tendentious: it would preclude, by definition, that administering an anti-prejudice pill would count as holding responsible. Usually, people are thought to be morally justified to perform actions of deliberative rationality as a means of influencing each other. Therefore, I decided to distinguish non-deliberative actions as a distinct class of actions and isolate manipulative actions as a species of the genus 'non-deliberative actions'. It gave me the conceptual room I needed to explore whether administering a pill or any other non-deliberative yet non-manipulative actions might count as holding responsible.

To focus my analysis, I made an epistemic turn. The best way to analyse how responsible agency is disturbed by hidden prejudices and possibly by non-deliberative actions, I assumed, is to treat it as an epistemic problem. It required two lines of reasoning. On the one hand, I needed to narrow down my analysis from agency, as an umbrella notion, towards *epistemic autonomy*, as a specific, epistemic condition of responsible agency.⁴ Exploring this notion would help us understand what the *epistemic* conditions for actions are and count as ones for which people are responsible. I distinguished two key features of epistemic autonomy: authorising and having self-knowledge (reflection). On the other hand, I needed to understand hidden prejudices and non-deliberative actions as epistemic phenomena, for example, by describing their influences on people in epistemic terms. Both lines of reasoning enabled an analysis of practices of hidden prejudice as rooted in a veil of prejudice: the boundary that divides people's self-knowledge and self-ignorance. Do people have (self)knowledge of their practices of prejudice *as* practices of prejudice? And: do people have knowledge of the effects of non-deliberative actions as the effects of these non-deliberative actions? Once we succeed in understanding responsible agency as epistemic autonomy and understanding hidden prejudices and non-deliberative *in epistemic terms*, we might also understand their interaction in epistemic terms and what this implies for *being* and *holding* responsible for practices of hidden prejudice.

Below I will summarise how the mentalist and pragmatist approaches understand the notions of epistemic autonomy, hidden prejudice and non-deliberative actions, second, how they conceptualise responsibility for practices of hidden prejudice, and, finally, how they answer the research question. I will start with the mentalist approach.

Mentalism understands the epistemic conditions of *agency* in terms of *mental states*, both conscious and unconscious, and those of *responsible agency* as *conscious, deliberative* and *individual*. From here, we can pursue two strategies. The first strategy is provided by the starting view: people are not responsible for their hidden prejudices. Epistemic autonomy is defined in terms of conscious, deliberative introspection. Hidden prejudices influence people's beliefs and actions in a way that bypasses their conscious, deliberative introspection.

Consequently, they cannot be held responsible. Here we get stuck in answering the research question. We do not arrive at the point at which we can answer the question 'under which conditions' because there is no 'holding responsible'. Another strategy would be to work out a view in which people are responsible for their prejudice. But as this cannot change the basic logic of 'responsible = conscious' and 'hidden prejudice = unconscious = bypassing introspection', this can only be an *indirect* responsibility. Even so, this might still open ways of holding people responsible, for example, by performing

⁴ Further analysis of responsible agency, of which epistemic agency is only one condition, might show that other conditions consist of certain forms of self-governing, for example, in a Frankfurtian sense.

non-deliberative actions. But as non-deliberative actions bypass introspection, by definition, it seems that they will always violate the norm of epistemic autonomy.

Pragmatism understands the epistemic conditions of responsible agency in terms of *rational, normative and social practices*. The strategy I have pursued is as follows. Epistemic autonomy turns on inferential (self-)knowledge, which should be understood as a continuum of more and less knowledgeable, of less and more ignorant, about one's actions. It means, for example, that *having hidden prejudices* is also, depending on the context, a mixture of knowledge and ignorance. This allows for claiming, again depending on the context, that people are responsible for their hidden prejudices in various degrees and in various forms. This brings us to the next question: would non-deliberative actions, which are aimed at reducing their hidden prejudice, violate their epistemic autonomy? If they did, they would not count as holding responsible. In my pragmatist analysis, I proposed that we understand epistemic autonomy as the freedom to disagree. Whatever non-deliberative intervention we choose to hold prejudiced people responsible, it should not violate their freedom to disagree. Manipulative actions offer a type of non-deliberative actions that violate this norm because, for such actions, it is essential that people do not recognise the effects of manipulation *as* the effects of manipulation. In discussing mentalism, I suggested administering an anti-prejudice pill as a non-deliberative action. From a pragmatist perspective, it might also count as a non-deliberative action. But I have a different reason for dismissing this type of action. From a pragmatist perspective practices, of hidden prejudice are social phenomena. Whatever interventions we can think of should therefore address these practices as social phenomena. An anti-prejudice pill fails to do just that. As an alternative, I proposed and explored *narrative performance* as an alternative type of non-deliberative action. Whether people have epistemic autonomy also depends on the question of whether a certain kind of epistemic balance exists between various social groups within in society: whether people from some social groups have, compared to people from other social groups, sufficient knowledge of how specific techniques or tools of non-deliberative action work (also see section 7.3.). In the case of my case study on news photographs, it implies that people have sufficient knowledge of how photography works, how it is used in the context of newspapers, et cetera, and of how the strategies of the civil rights movement work (their use of non-violent resistance). In my case study, I have assumed that the relevant audiences had sufficient knowledge of how these techniques work, compared to people from other social groups.⁵ People may disagree with this assumption. But one aim of my case study was only to illustrate how we may conceive of non-deliberative actions in social contexts. So, if people disagree with certain assumptions in my case study,

5 The relevant audiences at least had a familiarity with photographs. They had less familiarity with television which was, in those days, still a new medium. While the photographs of Elizabeth Eckford had not been manipulated, the television registration did. See, e.g. Roberts and Klibanoff (2006:160) on how a television crew noticed it had been too late in filming the screaming mob surrounding her and organised an 'artificial retake'.

they may think it is possible to find more plausible cases of people who have sufficient knowledge of how non-deliberative actions work. It is possible, theoretically at least and, as I think, in my case study as well, to think of non-deliberative actions that are not manipulative, i.e. do not violate the norm of the freedom to disagree. Therefore, what defines the epistemic conditions for non-deliberative actions to count as holding responsible, according to the pragmatist answer to the research question, is the *freedom to disagree*.

From this discussion, I conclude I achieved my research aim: the pragmatist perspective can make conceptual room for understanding non-deliberative actions as holding responsible; the mentalist perspective cannot.

8.3. Evaluating both frameworks

It might still be an open question whether achieving my research aim is of any significance. We might think of these two answers as simply two *alternative* answers to one research question, which cannot be compared for their strengths or weaknesses as they derive from perspectives that are based on two entirely different frameworks. After all, every notion that I proposed to answer the research question – responsible agency, epistemic autonomy, hidden prejudice, non-deliberative action – has a different meaning for each perspective. The choice between frameworks is one between paradigms. When you choose one framework, it means seeing different things than when you choose another.⁶ I nevertheless think that both frameworks can be compared and that their relative strengths can be assessed, for example, in terms of how nuanced their analyses are.

My analysis has tried to unravel the differences between the mentalist and pragmatist frameworks at a very basic level. One difference, which I proposed in the first chapter, is that the basic object of analysis for the mentalist framework is the thought process (thought as mental states that are either conscious or unconscious), while for the pragmatist framework, it is action. Another difference, which I discussed in chapters 4 and 5, is, very succinctly, that the pragmatist framework considers rationality and freedom (and its corresponding notion of responsibility) – or, to put it somewhat differently: knowledge (people as knowers) and agency (people as agents) – as closely connected while the mentalist framework treats rationality (in a broad, not just, deliberative, sense) and freedom/responsibility as analytically distinct features of being human.⁷ Instead, the mentalist framework connects freedom (and its corresponding notion of responsibility) only to a *reduced* notion of rationality: deliberative rationality. My analysis indicated that the mentalist framework has problematic implications

⁶ Johnson (2020:22), for example, refers to the mentalist methodology of psychology as a ‘paradigm’.

⁷ In section 1.4, I have briefly suggested that both frameworks start from two different approaches to agency: the mentalist from a ‘metaphysical’, the pragmatist from a ‘stance’ approach. Also, see Lance and White (2007).

for responsible agency. This is because it uses the conscious/unconscious distinction to conceptualise responsible agency *without* giving up on the assumption that rationality and freedom/responsibility are analytically distinct features. The logic of the dichotomous distinction between conscious and unconscious forces theorists to make the same dichotomous decisions, again and again, for other concepts. It results in aligning ‘conscious’ with for example ‘deliberative’, ‘slow’, ‘System 2’, ‘responsible’, ‘intentional’ and aligning ‘unconscious’ with ‘non-deliberative’, ‘fast’, ‘System 1’, ‘non-responsible’, ‘non-intentional’.⁸ Therefore, responsibility agency consists of conscious, intentional, deliberated, controlled choices for actions that are, to this extent, one’s own. But the notion of responsible agency that results from these alignments is a very unstable one. While ‘being deliberative’ is intimately connected to ‘being conscious’ – the two basic ingredients for ‘being responsible’ –, it is not intimately connected to ‘being transcendentally free’ as this is an analytically distinct feature. Consequently, the mental states that explain responsible actions – again: the conscious, deliberative, et cetera, ones – do not constitute a domain of entities that is immune to being explained in terms of unconscious states.⁹ Psychologists will continue to explain thought processes that qualify as ‘conscious’ and ‘deliberative’ in terms of thought processes that qualify as ‘unconscious’ and ‘non-deliberative’. The domain where people count as responsible agents (and therefore as agents acting from conscious/ deliberative reasons) is merely a transitory domain, a domain of actions whose roots in the human brain have not yet been sufficiently explained. In short: the mentalist notion of responsible agency is steadily dissolving as psychological and neurological research advances.

The strategy of the Brandomian-pragmatist framework was to distinguish between a transcendental and a moral-political level of analysis. At a metaphysical level, it argued for a notion of rationality that is *practical* (being rational is using concepts), *inferential* (using concepts is understanding inferential antecedents and inferential consequences: understanding when to use a concept and, if one uses a concept, understanding what will follow from that), *normative* (understanding a concept correctly means understanding that using this concept excludes the use of certain other concepts) and *social* (being rational is being rational in the company of others). These features combine to a development model of rationality – briefly: being rational is a matter of becoming rational, i.e. of

8 Some philosophers may disagree with this. They may claim, for example, that rationality can be unconscious (such views can be mostly found in the literature on dual-processes; see, e.g. Evans (2008) who distinguishes between ‘evolutionary rationality’, in the domain of System 1, and ‘individual rationality’, in the domain of System 2). But that would miss the point. Even in those cases that theorists acknowledge that not all rational processes need to be deliberative, the basic logic remains preserved: once it is claimed that rationality can be an unconscious process, that process at least cannot be deliberative (I do not know of any theorist who defends the claim that rational processes can be both unconscious and deliberative). And not giving up on the assumption that freedom and rationality are analytically distinct features is what allows mentalist theorists to explore the possibility of theorising rational processes as unconscious processes.

9 See Johnson (2020:23): ‘The hope is that we eventually arrive at an analysis constituted entirely by simple, elementary states’.

acquiring knowledge, of learning to use concepts. It means that knowledge/rationality and agency/freedom are tightly connected. The pragmatist transcendental argument was used to criticise the mentalist framework because it cannot explain the conditions that enable the acquisition of scientific knowledge or, broader, how people can develop their rational abilities.

The pragmatist criticism of the mentalist framework may not persuade all philosophers or social psychologists. It may seem to them that I have presented a simplified version of the mentalist framework. Philosophers and researchers on implicit bias sometimes recognise that our self-knowledge of implicit biases is not as rigid as earlier research has suggested.¹⁰ But these developments have not yet led to questioning or even exploring the mentalist framework itself. Many of them continue to work and to write as if their mentalist vocabulary is unproblematic. Some sociologists seem interested in adopting the ‘dual-process’ model for their theories, also accepting the underlying conscious/unconscious distinction.¹¹

My discussion at the metaphysical level of rationality and freedom may not have decided the case in favour of the pragmatist framework and against the mentalist framework. But perhaps we may find the tipping point once we evaluate both frameworks at the moral-political level by applying them to the urgent, social-political issue that started my research question: practices of prejudice. In this respect, my analysis has shown, I think, that the pragmatic approach offers various advantages over the mentalist framework, not only in describing or explaining the underlying moral-psychological features but also in suggesting what we still need to address in our analyses if we want to design new, successful strategies for reducing discrimination. I will briefly summarise my main criticisms:

1. *Individual versus social.* From a mentalist perspective, analysing the epistemic conditions of practices of hidden prejudices means identifying and assessing *individuals’* actions. From a pragmatist perspective, it means interpreting social practices in terms of (implicit) social intentionality, which is implicit and reproduced in various kinds of social practices. From a pragmatist perspective, the mentalist perspective can be criticised for ignoring why or how people, perhaps from certain social groups or in certain social positions, continue to have (hidden) prejudices. In short: the mentalist perspective narrows practices of discrimination down to individual actions of discrimination.

2. *Empirical versus normative.* To identify what is morally problematic with practices of hidden prejudice, it suffices, from a mentalist perspective, to describe the content of those implicit biases. Instead, the pragmatist perspective assumes that what is morally problematic with them depends on various norms that are shared in a community and

10 See, for example, Gawronski (2009:144), who addresses this issue by answering the question ‘Do Implicit Measures Provide a Window to the Unconscious?’. In his answer, he also discusses research findings that ‘are quite difficult to explain by the unconsciousness account’.

11 For an overview and discussion: see Leschziner (2019).

are open to debate, which are explicit in what is said and implicit in what is done. These norms include norms of non-discrimination and norms on having knowledge of the (possibly) discriminatory effects of one's actions and norms on what counts circumstances that exonerate individual responsibility. This perspective, therefore, requires not only to describe what people do know (either consciously or unconsciously), but also to assess whether, and to what extent, people ought (a) to have knowledge of norms of non-discrimination, (b) ought to have knowledge of the possibly discriminatory effects of their actions and (c) ought to acquire more knowledge about discrimination (norms and effects). Therefore, the pragmatist perspective might criticise the mentalist perspective for including the violation of norms of non-discrimination into the content of implicit biases. In short: the mentalist perspective narrows descriptions of morally problematic (prejudiced) practice down to descriptions of certain (supposedly) factual mental states.

3. *Unconscious versus ignorant.* Responsibility for practices of (hidden) prejudice means, from a mentalist perspective, that people have conscious knowledge of their actions as discriminatory actions. Therefore, the *assessment* of responsibility turns on the dichotomous application of a conscious/unconscious distinction, resulting in an equally dichotomous division between responsible and non-responsible (not: irresponsible) actions. For assessing responsibility for implicit bias, it means that, according to the starting view (section 7.8.), any one who has an implicit bias test result is not responsible for actions caused by her implicit biases. The pragmatist perspective understands *having hidden prejudices* on a continuum between more and less knowledgeable, between less and more ignorant, about one's actions: in terms of the *effects* of one's actions (performed as an individual, but also as belonging to certain social groups), also in terms of the *norms* for assessing actions and their effects. *Therefore, assessing* responsibility requires a *contextual* analysis: to analyse to what extent people had knowledge of their discriminatory actions *as* discriminatory actions, to what extent they could be expected to have certain knowledge (of norms or effects) and what might count as exonerating responsibility. From a pragmatist perspective, the mentalist perspective can be criticised for its rigid conscious/unconscious logic. It is at risk of subsuming cases of hidden prejudices under the heading of 'unconscious, therefore not responsible' that from a pragmatist perspective could have been analysed in terms of responsibility. Especially when research on implicit bias seems to show that hidden prejudice is a pervasive phenomenon in society, it raises the question to what extent its scientific insights, in their mentalist interpretation, are helpful to reduce practices of discrimination. In short: the mentalist perspective narrows the assessment of responsibility for hidden prejudice down to blunt simplifications.

4. *Administering an anti-prejudice pill versus narrative performance.* We might ignore how the mentalist perspective assesses moral responsibility for implicit biases, and we might assume that the mentalist perspective can somehow justify the administration of an anti-prejudice pill. From a mentalist perspective, it would still make sense to confront discrimination by administering an anti-prejudice pill: this might reduce

the negative emotions that prejudiced people feel about the person affected by their hidden prejudices. It assumes, as should be clear by now, that practices of prejudices are rooted in the unconscious, negative associations prejudice people have. However, from a pragmatist perspective, the ‘causes’ of practices of prejudices are multi-faceted. These ‘causes’ may refer to, for example, epistemic features – the extent to which people have knowledge, or are ignorant, of the facts and norms about discrimination –, to moral features – the extent to which people are committed to the value of non-discrimination –, to interests – the extent to which people, committed to non-discrimination, knowledgeable about the discriminative effects of their practices, still have an interest in not changing their practices –, to social practices of silencing or even to the products of these practices: laws, housing policies, psychological tests, et cetera. From a pragmatist perspective, the mentalist perspective may be criticised for designing *simplified* solutions – the administration of an anti-prejudice pill or similar interventions – for *complex* causes. It may be criticised, therefore, for designing solutions that will be relatively unsuccessful in confronting discrimination. What can the pragmatist perspective offer as an alternative to the anti-prejudice pill? I have argued that the pragmatist perspective offers conceptual room for designing non-deliberative yet non-manipulative interventions for which the mentalist perspective does not offer any room. Illustrative for such pragmatist interventions are *narrative performances*. While these have an equally hard job in confronting discrimination effectively, in the long term, they hold the promise for unravelling, thread by thread, the veil of prejudice. But I concede that this is also a matter for further research (see next section).

Comparing the mentalist and pragmatist frameworks in how they analyse, at the moral-political level, the practices of hidden prejudice, has made it plausible, at a minimum, that the pragmatist framework provides an alternative, consistent analysis of practices of hidden prejudice and of responsibility for them. But I think comparing them also has made it plausible that the pragmatist perspective is able to provide a more fine-grained analysis than the mentalist perspective: it allows for the possibility to analyse practices, not just of one individual, but as social practices (in various senses of ‘social’), as practices that are also practices of norms, finally, as practices for which people have a responsibility in various ways.

The comparison at the moral-political level cannot settle the comparison on the metaphysical level. Nevertheless, I would like to make one last suggestion for rejecting the mentalist framework. On a metaphysical level, the mentalist strategy for both conceptualising *and* naturalising responsible agency turns on, amongst others, the dichotomy between conscious and unconscious. To the extent that this dichotomy does not hold, this strategy is undermined. In the end, the discussion on the unconsciousness of implicit biases turns on the question of whether we accept the conscious/unconscious distinction as the founding distinction, which also underlies and informs parallel distinctions such as rational/non-rational, responsible/non-responsible. My pragmatist

analysis was intended to put into doubt the claim that (hidden) prejudices can be described in a dichotomous, simplifying vocabulary of conscious/unconscious. And it proposed a more nuanced assessment of responsibility. If my analysis has persuaded any of my readers, I suggest they should start considering the rejection of the underlying mentalist framework.

8.4. Suggestions for future research

In different ways, I have tried to push my research as far as my research question would allow me. But I have not pursued and explored all the possible lines of argument that might have bolstered my analysis. As a result, my argument still suffers from various gaps and weaknesses. Below, I will suggest future research from three different perspectives: a historical, methodological, and thematic perspective.

Historical perspective. In my research, I have not necessarily assumed that liberal *political* philosophers are committed to a mentalist framework. Instead, I have chosen a weak claim: their notions of autonomy and deliberative rationality *can* be assimilated into a mentalist framework. The strong claim would be that liberal philosophers, at least those committed to notions of autonomy and deliberative rationality, are committed to a mentalist framework. Defending this claim requires a detailed argument that would have been too far beyond the scope of my research.¹² Nevertheless, I think it would be constructive to investigate more systematically, from a historical perspective, whether and to what extent liberal, political philosophers took notice and used insights from psychological research.

Methodological perspective. While I have frequently suggested that assessing responsibility for hidden prejudice requires a contextual approach from a pragmatist perspective, I have not shown how this works in particular cases. This is really a gap in my argument. But it is also suitable for future research. Working out a pragmatist methodology for analysing cases of prejudice would also require, I think, a more developed theory of responsibility. I have assumed, rather than made explicit, that holding people responsible may cover a wide range of different actions which do not need to have a punitive function but perhaps may have a dialogical or conciliatory function. Furthermore, it would be interesting to explore the conceptual links between

12 I think it would require, amongst others, a historical reconstruction of how Anglo-American liberal philosophers relied for their notions of agency (autonomy) on psychological ideas or how they responded more generally to developments in psychology. Such studies seem to be scarce. But see, e.g. Daston (1978), who describes how (amongst others) late nineteenth-century British philosophers responded to the new psychological ideas. Historical overviews of 'introspection' in philosophy at least suggest that 'introspection' was important to the Anglo-American tradition of philosophy while it was much less important, for example, to the German philosophical tradition (also, see chapter 2, section 6). I think that Frankfurt (1971) provides an important key to unravelling liberalism's commitment to the mentalist framework. More about this in chapter 2, section 3, although I will not provide any detailed argument.

pragmatism and phenomenology in more detail, especially because both are used for developing social notions of habit.¹³

Thematic perspective. Various thematic issues require more detailed analyses, but one of them would certainly be: *non-deliberative action as social-political action*. I have tried to conceptualise it as narrative performance and discussed some cases of news photographs. But this analysis was still too underdeveloped. I think this should be further examined. For example, the idea of non-deliberative action as narrative performance suggests conceptual links with rhetoric: although rhetorical analysis has traditionally focussed on speech, it has often been interested in its non-deliberative dimension to influence people. Exploring these links might help to further develop the notion of narrative performance. And in whatever way this notion is further explored, it should again be closely connected to a case study.

As a final note, I want to remind my readers that I pursued my research in the context of liberalism as a political philosophy and of how scientific insights into responsible agency inform it. In my view, it provides a powerful source for African Americans, Roma people and many other groups of people to develop and articulate emancipatory appeals for equality. For these appeals and the efforts of politicians, institutions, and ordinary citizens, to be effective in confronting discrimination practices, scientific insights may prove to be valuable and essential. At the same time, the lesson that I hope we might learn from my analysis is that assumptions at a deep theoretical level, one that I have called 'framework', already steer our understanding – of responsible agency, of ways to confront discrimination – in certain, sometimes unhelpful, directions. I recommend a pragmatist framework.

13 See, e.g. Ngo (2017).

Appendices

Bibliography

Summary

Samenvatting in het Nederlands

Curriculum Vitae

Quaestiones Ininitae

Bibliography

Dates of books or articles will usually refer to the original publication. Some books and articles are listed with two dates of publication: the first referring to the date of the original publication, the second referring to the date of publication I used.

- Abbey, Ruth and Spinner-Halev, Jeff (2013) 'Rawls, Mill, and the Puzzle of Political Liberalism' *The Journal of Politics*, 75:1, 124–136.
- Abel, Elizabeth (2010) *Signs of the Times. The Visual Politics of Jim Crow*. University of California Press.
- Adams, Glenn; Monica Biernat; Nyla R. Branscome; Christian S. Crandall; Lawrence S. Wrightsman, eds. (2008a) *Commemorating Brown. The Social Psychology of Racism and Discrimination*. American Psychological Association.
- (2008b) 'Beyond Prejudice: Toward a Sociocultural Psychology of Racism and Oppression' in Adams et al. (2008a), Chapter 12, 215–246.
- Adamson, Peter (2014) 'Freedom and determinism' in Pasnau and Van Dyke (2014), Chapter 29, 399–413.
- Ahmed, Leila (1992) *Women and Gender in Islam. Historical Roots of a Modern Debate*. Yale University Press.
- Albert, Hans (1968/ 1991) *Traktat über kritische Vernunft*. Mohr Siebeck.
- Alcoff, Linda and Potter, Elizabeth, eds. (1993) *Feminist Epistemologies*. Routledge.
- Alexander, Michelle (2010/ 2012) *The New Jim Crow. Mass Incarceration in the Age of Colorblindness*. The New Press.
- Allen, Danielle S. (2004) *Talking to strangers. Anxieties of Citizenship since Brown v. Board of Education*. The University of Chicago Press.
- Allison Henry E. (1990) *Kant's Theory of Freedom*. Cambridge University Press.
- Allport, Gordon W. (1954/ 1979) *The Nature of Prejudice. 25th Anniversary Edition*. Basic Books.
- Alston, William P. (1986) 'Internalism and Externalism in Epistemology' *Philosophical Topics*, 14:1, Papers on Epistemology, 179–221.
- (1988) 'The Deontological Conception of Epistemic Justification' *Philosophical Perspectives*, 2, 257–299.
- Altman, Neil and Johanna Tiemann (2004) 'Racism as Manic Defense' in Levine and Pataki (2004), 127–141.
- Ameriks, Karl, ed. (2000) *The Cambridge Companion to German Idealism*. Cambridge University Press.
- Amnesty International; European Roma Rights Centre; Open Society Fund and Czech Society for Inclusive Education (2015) *Czech Republic: Eight years after the D.H. judgement a comprehensive desegregation of schools must take place*. Public statement published on 14 November 2015. Index: EUR 71/2863/2015. Available at: <https://>

- www.amnesty.org/download/Documents/EUR7128632015ENGLISH.pdf (accessed on 25 March 2021).
- Amodio, David M. and Mendoza, Saaid A. (2010) 'Implicit Intergroup Bias. Cognitive, Affective, and Motivational Underpinnings' in Gawronski and Payne (2010), Chapter 19, 353–374.
- Amsterdam (municipality of) (2019) 'Werkplan Aanpak Arbeidsmarktdiscriminatie 2019-2020' <https://www.binnenlandsbestuur.nl/Uploads/2019/7/werkplan-aanpak-arbeidsmarktdiscriminatie-def-1.pdf> (accessed on 4 September 2020).
- Anderson, Benedict (1983) *Imagined Communities*. Verso.
- Anderson, Scott A. (2010) 'How Did There Come To Be Two Kinds of Coercion?' in Reidy and Riker (2010), 17–29.
- Anderson, Joel (2014) 'Regimes of Autonomy' *Ethical Theory and Moral Practice*, 17, 355–368.
- Anderson, Michael L. and Perlis, Don (2009) 'What puts the "meta" in metacognition?' *Behavioral and Brain Sciences*, 32:2, 18–19.
- Andriopoulos, Stefan (2000) *Besessene Körper. Hypnose, Körperschaften und die Erfindung des Kinos*. Wilhelm Fink Verlag.
- (2011) 'The Sleeper Effect: Hypnotism, Mind Control, Terrorism' Grey Room, 45, 88–105.
- Anscombe, G.E.M. (1957/ 1963) *Intention*. Second Edition. Cornell University Press.
- Archard, David (1992) 'Autonomy, character and situation' in Milligan and Miller (1992), 157–170.
- Arendt, Hannah (1957) 'Reflections on Little Rock' in Arendt (2003b), 193–213.
- (2003a) 'Collective Responsibility' in Arendt (2003b), 147–158.
- (2003b) *Responsibility and Judgment*. Schocken Books.
- (2006a) *Between Past and Future*. Penguin Books.
- (2006b) 'What is Freedom?' in Arendt (2006a), Chapter 4, 142–169.
- Armour, Jody (1995) 'Stereotypes and Prejudice: Helping Legal Decisionmakers Break the Prejudice Habit' *California Law Review*, 83:3, 733–772.
- Armour–Garb, Bradley P. and JC Beall (2005a) 'Deflationism: The Basics' in Armour–Garb and Beall (2005b), 1–29.
- ,eds. (2005b) *Deflationary Truth*. Open Court Publishing Company.
- Arpaly, Nomy (2003) *Unprincipled Virtue. An Inquiry into Moral Agency*. Oxford University Press.
- (2004) 'Which Autonomy?' in Campbell, O'Rourke and Shier (2004), Chapter 8, 173–188.
- (2006) *Merit, Meaning, and Human Bondage. An Essay on Free Will*. Princeton University Press.
- Arrington, Robert L. (1982) 'Advertising and behavior control' *Journal of Business Ethics*, 1:1, 3–12.

- Atmanspracher, Harald and Bishop, Robert, eds. (2002) *Between Chance and Choice. Interdisciplinary Perspectives on Determinism*. Imprint Academic.
- Audi, Robert (1991) 'Responsible Action and Virtuous Character' *Ethics*, 101:2, 304–321.
- Ausborn-Brinker (1999) *Person und Personalität. Versuch einer Begriffsklärung*. Mohr Siebeck.
- Ayer, A.J. (1954) 'Freedom and Necessity' in Watson (1982), 15–23.
- (1963a) 'The Concept of a Person' in Ayer (1963b), 82–128.
- (1963b) *The Concept of a Person and other Essays*. MacMillan & Co LTD.
- Bader, Veit (1988) 'Macht of waarheid? De 'kennisociologische blindheid' van Kunnemans communicatietheoretisch perspectief en de 'normatieve blindheid' van Pels' rivaliteitsperspectief op wetenschap' *Kennis en Methode*, XII:2, 138–157.
- (1995) *Rassismus, Ethnizität, Bürgerschaft. Soziologische und philosophische Überlegungen*. Westfälisches Dampfboot.
- (1991) *Collectief Handelen. Sociale ongelijkheid en collectief handelen. Deel 2*. Wolters-Noordhoff.
- (2001) 'Problems and Prospects of Associative Democracy' in Hirst and Bader (2001), 31–70.
- (2005) 'Reasonable Impartiality and Priority for Compatriots. A Criticism of Liberal Nationalism's Main Flaws' *Ethical Theory and Moral Practice*, 8, 83–103.
- (2007a) *Secularism or Democracy? Associational Governance of Religious Diversity*. Amsterdam University Press.
- (2007b) 'Misrecognition, Power, and Democracy' in van den Brink and Owen (2007), 238–269.
- (2013) 'Free Speech or Non-discrimination as Trump? Reflections on Contextualised Reasonable Balancing and Its Limits' *Journal of Ethnic and Migration Studies*, DOI: 10.1080/1369183X.2013.851478.
- Bader, Veit-Michael and Albert Benschop (1988) *Ongelijkheden. Sociale ongelijkheid en collectief handelen. Deel 1*. Wolters-Noordhoff.
- Baer, John; Kaufman, James C. and Baumeister, Roy F., eds (2008) *Are We Free? Psychology and Free Will*. Oxford University Press.
- Banaji, Mahzarin R.; Bhaskar, Roy and Brownstein, Michael (2015) 'When bias is implicit, how might we think about repairing harm?' *Current Opinion in Psychology*, 6, 183–188.
- Banks, Ralph Richard, and Richard Thompson Ford (2009) '(How) Does Unconscious Bias Matter: law, Politics, and Racial Inequality' *Emory Law Journal*, 58:5, 1053–1122.
- Baker, Lynne Rudder (1998) 'The First-Person Perspective: A Test for Naturalism' *American Philosophical Quarterly*, 35:4, 327–348.
- (2013) *Naturalism and the First-Person Perspective*. Oxford University Press.

- Barbalet, J.M. (1985) 'Power and Resistance' *The British Journal of Sociology*, 36:4, 531–548.
- Barber, Michael D. (2011) *The Intentional Spectrum and Intersubjectivity. Phenomenology and the Pittsburgh Neo-Hegelians*. Ohio University Press.
- Barclay, Linda (2000) 'Autonomy and the Social Self' in Mackenzie and Stoljar (2000), 52–71.
- Bargh, John A. (1990) 'Goal ≠ Intent: Goal-Directed Thought and Behavior Are Often Unintentional' *Psychological Inquiry*, 1:3, 248–251.
- (1994) 'The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition' in Wyer and Srull (1994), Chapter 1, 1–40.
- (1999) The cognitive monster: The case against the controllability of automatic stereotype effects in S. Chaiken and Y. Trope (1999), Chapter 18, 361–382.
- ,ed. (2007) *Social Psychology and the Unconscious. The Automaticity of Higher Mental Processes*. Psychology Press.
- (2008) 'Free Will is Un-natural' in Baer, Kaufman and Baumeister (2008), Chapter 7, 128–154.
- Bargh, John A. and Melissa J. Ferguson (2000) 'Beyond Behaviorism: On the Automaticity of Higher Mental Processes' *Psychological Bulletin*, 126:6, 925–945.
- Bargh, John A. and Ezequiel Morsella (2008) 'The Unconscious Mind' *Perspectives on Psychological Science*, 3:1, From Philosophical Thinking to Psychological Empiricism, Part I, 73–79.
- Bargh, John A. and Brian D. Earp (2009) 'The Will is Caused, not 'Free'' *Dialogue: Newsletter of the Society for Personality and Social Psychology*, 24:1, 13–15.
- Barnett, Clive (2017) *The Priority of Injustice. Locating Democracy in Critical Theory*. The University of Georgia Press.
- Barnhill, Anne (2014) 'What Is Manipulation?' in Coons and Weber (2014), 51–72.
- Bar-Tal, Daniel; Graumann, Carl F.; Kruglanski, Arie W. and Stroebe, Wolfgang, eds. (1989) *Stereotyping and Prejudice. Changing Conceptions*. Springer-Verlag.
- Basker, James G., ed. (2012) *American Antislavery Writings. Colonial Beginnings to Emancipation*. The Library of America.
- Baston, René (2020) *Implizite Vorurteile. Wie unbewusster Rassismus unser Denken begleitet*. J.B. Metzler.
- Baum, Bruce and Robert Nichols, eds. (2013) *Isaiah Berlin and the Politics of Freedom. "Two Concepts of Liberty" 50 Years Later*. Routledge.
- Baumgold, Deborah (2010) *Contract Theory in Historical Context. Essays on Grotius, Hobbes, and Locke*. Brill.
- Bayertz, Kurt (1995a) 'Eine kurze Geschichte der Herkunft der Verantwortung' in Bayertz (1995b), 3–71.
- ,ed. (1995b) *Verantwortung. Prinzip oder Problem?* Wissenschaftliche Buchgesellschaft.

- Beals, M.B. (1994/ 2007) *Warriors Don't Cry: Searing Memoir of Battle to Integrate Little Rock*. Washington Square Press.
- Beauchamp, Tom L. (2005) 'Who Deserves Autonomy, and Whose Autonomy Deserves Respect?' in James Stacey Taylor, ed. (2005) *Personal Autonomy. New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, 1–29. Cambridge University Press.
- Beck, Sam and Ana Ivasiuc, eds. (2018) *Roma Activism. Reimagining Power and Knowledge*. Berghahnbooks.
- Bedau, Hugo Adam (2002) 'Compensatory Justice and the Black Manifesto' in Roberts (2002), 131–146.
- Beeghly, Erin and Madva, Alex (2020) *An Introduction to Implicit Bias. Knowledge, Justice, and the Social Mind*. Routledge.
- Beetz, Manfred (1983) 'Transparent gemachte Vorurteile. Zur Analyse der *praejudicia auctoritatis et praecipitantiæ* in der Frühaufklärung' *Rhetorik*, 3, 7–33.
- Beisecker, David (2006) 'Dennett's Overlooked Originality' *Minds and Machines*, 16:1, 43–55.
- Beiser, Frederick C. (1996) *The Sovereignty of Reason: The Defense of Rationality in the Early English Enlightenment*. Princeton, New Jersey: Princeton University Press.
- Bell, Daniel (1993) *Communitarianism and its Critics*. Clarendon Press.
- Bellamy, Richard (2000) *Rethinking Liberalism*. Pinter.
- Belton, Robert (2014) *The Crusade for Equality in the Workplace. The Griggs v. Duke Power Story*. University Press of Kansas.
- Benn, Stanley I. (1967) 'Freedom and Persuasion' *Australasian Journal of Philosophy*, 45:3, 259–275.
- (1975 – 1976) 'Freedom, Autonomy and the Concept of a Person' *Proceedings of the Aristotelian Society*, New Series, 76, 109–130.
- Benn, Stanley I. (1971) 'Being Free to Act, and Being a Free Man' *Mind*, New Series, 80:318, 194–211.
- Bennett, Jonathan (1974) 'The Conscience of Huckleberry Finn' *Philosophy*, 49:188, 123–134.
- Bergelson, Vera (1996–1997) 'Essay. Crimes and Defenses of Rodion Raskolnikov' *Kentucky Law Journal*, 85, 919–953.
- Berger, Martin A. (2011) *Seeing through race. A reinterpretation of civil rights photography*. University of California Press.
- (2013) *Freedom Now! Forgotten Photographs of the Civil Rights Struggle*. University of California Press.
- Berger, Peter, and Luckmann, Thomas (1989) *The Social Construction of Reality. A Treatise in the Sociology of Knowledge*. Anchor Books.
- Bergh, Linda van den; Eddie Denessen; Lisette Hornstra, Marinus Voeten and Rob W. Holland (2010) 'The Implicit Prejudiced Attitudes of Teachers: Relations to

- Teacher Expectations and the Ethnic Achievement Gap’.
- Berlin, Isaiah (1954) ‘Historical Inevitability’ in Hardy (2002), 94–165.
- (1958) ‘Two Concepts of Liberty’ in Hardy (2002), 166–217.
- (1963–1964) “From Hope and Fear Set Free” *Proceedings of the Aristotelian Society*, New Series, 64, 1–30.
- (1969/ 2002) ‘Introduction’ in Hardy (2002), 3–54.
- Bernasconi, Robert (2008) ‘Before Whom and What? Accountability and the Invention of Ministerial, Hyperbolic, and Infinite Responsibility’ in Sullivan and Schmidt (2008), 131–146.
- Berofsky, Bernard, ed. (1966) *Free Will and Determinism*. Harper & Row.
- (1995) *Liberation from Self. A Theory of Personal Autonomy*. Cambridge University Press.
- Bhaskar, Roy (1975) *A Realist Theory of Science*. Verso.
- Billig, Michael (2012) ‘The notion of ‘prejudice’: some rhetorical and ideological aspects’ in Dixon and Levine (2012), 139–157.
- Blaakman, Martin (2012) ‘Civil Disobedience in a Distorted Public Sphere’ *Krisis*, 3, 27–36.
- Black, Antony (1993) ‘The Juristic Origins of Social Contract Theory’ *History of Political Thought*, XIV:1, 57–76.
- Block, Ned (1995) ‘On a confusion about a function of consciousness’. *Behavioral and Brain Sciences*, 18, 227–287.
- Bloor, David (1997) *Wittgenstein, Rules and Institutions*. Routledge.
- Blöser, Claudia (2014) *Zurechnung bei Kant. Zum Zusammenhang von Person und Handlung in Kants praktischer Philosophie*. De Gruyter.
- Blum, Lawrence (2002) “I’m Not a Racist, But...” *The Moral Quandary of Race*. Cornell University Press.
- (2004) ‘Stereotypes And Stereotyping: A Moral Analysis’ *Philosophical Papers*, 33:3, 251–289.
- Bodroghkozy, Aniko (2012) *Equal Time. Television and the Civil Rights Movement*. University of Illinois Press.
- Bok, Hilary (1998) *Freedom and Responsibility*. Princeton University Press.
- Bolling v. Sharpe*, 347 U.S. 497 (1954).
- Bonilla-Silva, Eduardo; Lewis, Amanda and Embrick, David G. (2004) “‘I Did Not Get that Job Because of a Black Man...’: The Story Lines and Testimonies of Color-Blind Racism’ *Sociological Forum*, 19:4, 555–581.
- Borchert, Donald M., ed. (2006) *Encyclopedia of Philosophy. Volume 3 (Determinables – Fuzzy Logic)*. Second Edition. Thomson Gale.
- Borges, Jorge Luis (1944) ‘Funes el memorioso’ in Borges (2006), 519–525.
- (2006) *Obras Completas, 1, 1923 – 1949*. Emecé Editores.
- Boucher, David and Paul Kelly, eds. (1994) *The Social Contract from Hobbes to Rawls*.

- Routledge.
- (1994) 'The social contract and its critics. An overview' in Boucher and Kelly (1994), 1–34.
- Bouchilloux, Hélène (1997) 'Grotius et la question du droit de mentir' in Foisneau (1997), 131–154.
- Bourdieu, Pierre (1972/ 1977) *Outline of a Theory of Practice*. Cambridge University Press.
- (1980/ 1990) *The Logic of Practice*.
- (1990) *In Other Words. Essays Towards a Reflexive Sociology*. Stanford University Press.
- (1991) *Language & Symbolic Power*. Polity Press.
- Bourdieu, Pierre and Eagleton, Terry (1992) 'Doxa and Common Life' *New Left Review*, Issue 191, 111–121.
- Bourke, John (1938) 'Responsibility, Freedom and Determinism'. *Philosophy*, 13:51, 276–287.
- Bowers, L. (2003) 'Manipulation: searching for an understanding' *Journal of Psychiatric and Mental Health Nursing*, 10, 329–334.
- Bowers, K.S. and Meichenbaum, D., eds. (1984) *The Unconscious Reconsidered*. Wiley.
- Bradley, F.H. (1876/ 2006) *Ethical Studies*. The Clarendon Press.
- Brady, Michael S. and Fricker, Miranda, eds. (2016) *The Epistemic Life of Groups. Essays in the Epistemology of Collectives*. Oxford University Press.
- Brandom, Robert. (1979) 'Freedom and Constraint by Norms' *American Philosophical Quarterly*, 16:3, 187–196.
- (1994) *Making It Explicit. Reasoning, Representing, and Discursive Commitment*. Harvard University Press.
- (1997) 'Replies' *Philosophy and Phenomenological Research*, 57:1, 189–204.
- (1998) 'Insights and Blindspots of Reliabilism' *The Monist*, 81:3, 371–392.
- (1999) 'Some Pragmatist Themes in Hegel's Idealism: Negotiation and Administration in Hegel's Account of the Structure and Content of Conceptual Norms' *European Journal of Philosophy*, 7:2, 164 – 189.
- (2000a) *Articulating Reasons. An Introduction to Inferentialism*. Harvard University Press.
- (2000b) 'Facts, Norms, and Normative Facts: A Reply to Habermas' *European Journal of Philosophy*, 8:3, 356 – 374.
- (2001a) 'Modality, Normativity, and Intentionality' *Philosophy and Phenomenological Research*, 63:3, 587–609.
- (2001b) 'Action, Norms, and Practical Reasoning' in Millgram (2001), Chapter 20, 465–479.
- (2001c) 'What do Expressions of Preferences Express?' in Morris and Ripstein (2001), 11–36.

- (2002a) *Tales of the Mighty Dead. Historical Essays in the History of Intentionality*. Harvard University Press.
- (2002b) ‘Facts, Norms, and Normative Facts: A Reply to Habermas’ *European Journal of Philosophy*, 8:3, 356–374.
- (2004a) ‘The Pragmatist Enlightenment (and its Problematic Semantics)’ *European Journal of Philosophy*, 12:1, 1–16.
- (2004b) ‘From a critique of cognitive internalism to a conception of objective spirit: reflections on Descombes’ Anthropological Holism’ *Inquiry: An Interdisciplinary Journal of Philosophy*, 47:3, 236–253, DOI: 10.1080/00201740410006357.
- (2005) ‘Responses to Pippin, Macbeth and Haugeland’ *European Journal of Philosophy*, 13:3, 429–441.
- (2006) ‘Kantian Lessons about Mind, Meaning, and Rationality’ *The Southern Journal of Philosophy*, XLIV: 49–71.
- (2007) ‘The structure of desire and recognition: Self-consciousness and self-constitution’ *Philosophy & Social Criticism*, 33:1, 127–150.
- (2008a) *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford University Press.
- (2008b) Reply to “Of Mu-Mesons and Oranges” Apel, Bahrenberg, Köhne, Prien, Suhm’ in Prien and Schweikard (2008), 167–171.
- (2009) *Reason in Philosophy*. The Belknap Press of Harvard University Press.
- (2011) ‘Platforms, Patchworks, and Parking Garages: Wilson’s Account of Conceptual Fine-Structure in Wandering Significance’ *Philosophy and Phenomenological Research*, LXXXII:1, 183–201.
- (2015) *From Empiricism to Expressivism. Brandom reads Sellars*. Harvard University Press.
- Brasil-Neto, Joaquim P.; Pascual-Leone, Alvaro; Valls-Solé, Josep; Cohen, Leonardo G. and Hallett, Mark (1992) ‘Focal transcranial magnetic stimulation and response bias in a forced-choice task’ *Journal of Neurology, Neurosurgery, and Psychiatry*, 55: 964–966.
- Bratman, Michael (1992) ‘Shared Cooperative Activity’ *The Philosophical Review*, 101:2, 327–341.
- (2003) ‘Autonomy and Hierarchy’ in Paul, Miller and Paul (2003), 156–176.
- Brekhus, Wayne H. and Gabe Ignatow (2019) *The Oxford Handbook of Cognitive Sociology*. Oxford University Press.
- Brennan, Samantha (2013) ‘Rethinking the Moral Significance of Micro-Inequities’ in Hutchison and Jenkins (2013), Chapter 9, 180–196.
- (2016) ‘The Moral Status of Micro-Inequities. In Favor of Institutional Solutions’ in Brownstein and Saul (2016), Chapter 3.3, 235–253.
- Brett, Nathan (1981) ‘Human Habits’ *Canadian Journal of Philosophy*, XI:3, 357–376.
- Brief, Arthur P.; Joerg Dietz; Robin Reizenstein Cohen; S. Douglas Pugh and Joel B.

- Vaslow (2000) 'Just Doing Business: Modern Racism and Obedience to Authority as Explanations for Employment Discrimination' *Organizational Behavior and Human Decision Processes*, 81:1, 72–97.
- Brink, Bert van den (2005) 'Liberalism without Agreement: Political Autonomy and Agonistic Citizenship' in Christman and Anderson (2005), 245 – 271.
- Brink, Bert van den and Owen, David, eds. (2007) *Recognition and Power. Axel Honneth and the Tradition of Critical Social Theory*. Cambridge University Press.
- Brock, Adrian C. (2013) 'The History of Introspection Revisited' in Clegg (2013), Chapter 2, 25–43.
- Bromand, Joachim and Guido Kreis, eds. (2010) *Was sich nicht sagen lässt. Das Nicht-Begriffliche in Wissenschaft, Kunst und Religion*. Akademie Verlag.
- Brown, Rupert (2010) *Prejudice. Its Social Psychology*. Second Edition. Wiley-Blackwell.
- Brown, Stuart, ed. (1996) *British Philosophy and the Age of Enlightenment. Routledge History of Philosophy. Volume 5*. Routledge.
- Brownstein, Michael (2015) 'Attributionism and Moral Responsibility for Implicit Bias' *Review of Philosophy and Psychology*, 1–22.
- (2019) 'Implicit Bias' in *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/>>.
- Brownstein, Michael and Jennifer Saul, eds. (2016) *Implicit Bias and Philosophy. Volume 2: Moral Responsibility, Structural Injustice, and Ethics*. Oxford University Press.
- Bruner, Jerome (1992) 'Another Look at New Look 1'. *American Psychologist*, 47:6, 780–783.
- Buckle, Stephen (1991) *Natural Law and the Theory of Property: Grotius to Hume*. Clarendon Press.
- Busche, Hubertus, ed. (2011) *Departure for Modern Europe. A Handbook of Early Modern Philosophy (1400 – 1700)*. Felix Meiner Verlag.
- Buss, Sarah (2012) 'Autonomous Action: Self-Determination in the Passive Mode' *Ethics*, 122:4, 647–691.
- Cairns, John W. (2001 – 2002) 'Stoicism, slavery, and law. Grotian Jurisprudence and Its Reception' *Grotiana, New Series*, 22/23, 197–232.
- Camic, Charles (1986) 'The Matter of Habit' *American Journal of Sociology*, 91, 1039–1087.
- Campbell, Joseph Keim; O'Rourke, Michael and Shier, David (2004) 'Freedom and Determinism: A Framework' in Campbell, O'Rourke and Shier (2004), 1–17.
- ,eds. (2004) *Freedom and Determinism*. The MIT Press.
- Cantor, G.N. (1975) 'The Edinburgh Phrenology Debate: 1803–1828' *Annals of Science*, 32:3, 195 – 218.
- Carmichael, Stokely and Hamilton, Charles V. (1967) *Black Power. The Politics of Liberation in America*. Jonathan Cape.

- Carruthers, Peter (1986) *Introducing Persons. Theories and Arguments in the Philosophy of Mind*. Routledge.
- (1992) *Human Knowledge and Human Nature. A new introduction to an ancient debate*. Oxford University Press.
- (1996a) ‘Simulation and self-knowledge: a defence of theory-theory’ in Carruthers and Smith (1996), Chapter 3, 22–38.
- (1996b) *Language, thought and consciousness. An essay in philosophical psychology*. Cambridge University Press.
- (2000) *Phenomenal consciousness. A naturalistic theory*. Cambridge University Press.
- (2006) *The Architecture of the Mind*. Clarendon Press.
- (2007) ‘The illusion of conscious will’ *Synthese*, 159, 197–213.
- (2009a) ‘An architecture for dual reasoning’ in Evans and Frankish (2009), Chapter 5, 109–127.
- (2009b) ‘How we know our own minds: The relationship between mindreading and metacognition’. *Behavioral and Brain Sciences*, 32:2, 1–18.
- (2010) ‘Introspection: Divided and Partly Eliminated’ *Philosophy and Phenomenological Research*, LXXX:1, 76–111.
- (2011) *The Opacity of Mind. An Integrative Theory of Self-Knowledge*. Oxford University Press.
- (2013a) ‘On Knowing Your Own Beliefs: A Representationalist Account’ in Nottelmann (2013), Chapter 7, 145–165.
- (2013b) ‘Animal minds are real, (distinctively human minds are not)’ *American Philosophical Quarterly*, 50:3, 233–248.
- (2014) ‘On central cognition’ *Philosophical Studies*, 170:1, 143–162.
- (2015) *The Centered Mind. What the Science of Working Memory Shows Us About the Nature of Human Thought*. Oxford University Press.
- (2016) ‘Higher-Order Theories of Consciousness’ in *The Stanford Encyclopedia of Philosophy* (Fall 2016 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/fall2016/entries/consciousness-higher/>>.
- Carruthers, Peter and Peter K. Smith, eds. (1996) *Theories of theories of mind*. Cambridge University Press.
- Carruthers, Susan L. (1998) ‘“The Manchurian Candidate’ (1962) and the Cold War brainwashing scare’ *Historical Journal of Film, Radio and Television*, 18:1, 75–94.
- Caruso, Gregg D. (2012) *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lexington Books.
- Cassirer, Ernst (1936/ 2004) *Determinismus und Indeterminismus in der modernen Physik. Historische und systematische Studien zum Kausalproblem. Gesammelte Werke Hamburger Ausgabe, Band 19*. Wissenschaftliche Buchgesellschaft.
- Celikates, Robin (2016) ‘Rethinking Civil Disobedience as a Practice of Contestation – Beyond the Liberal Paradigm’ *Constellations*, 23:1, 37–45.

- Chaiken, S. and Y. Trope, eds. (1999) *Dual-process theories in social psychology*. Guilford Press.
- Chapman, Michael (1990) 'Intention, Intentionality, and the Constructive Character of Scientific Knowledge' *Psychological Inquiry*, 1:3, 252 – 253.
- Chappell, Vere, ed. (1999) *Hobbes and Bramhall on Liberty and Necessity*. Cambridge University Press.
- Chetty, Ray; Nathaniel Hendren; Maggie R. Jones and Sonya R. Porter (2020) 'Race and Economic Opportunity in the United States: An Intergenerational Perspective' *The Quarterly Journal of Economics*, 135:2, 711–783.
- Christman, John (1987) 'Autonomy: A Defense of the Split-Level Self' *The Southern Journal of Philosophy*, XXV:3, 281–293.
- (1988) 'Constructing the Inner Citadel: Recent Work on the Concept of Autonomy' in *Ethics*, 99:1, 109–124.
- (1989) 'Introduction' in John Christman, ed. (1989) *The Inner Citadel. Essays on Individual Autonomy*, 3–23. Oxford University Press.
- (1991) 'Liberalism and Individual Positive Freedom' *Ethics*, 101:2, 343–359.
- (2005a) 'Saving Positive Freedom' *Political Theory*, 33, 79–88.
- (2005b) 'Autonomy, Self-Knowledge, and Liberal Legitimacy' in Christman, John and Joel Anderson, eds. (2005) *Autonomy and the Challenges to Liberalism. New Essays*. Cambridge University Press.
- (2009) *The Politics of Persons. Individual Autonomy and Socio-historical Selves*. Cambridge University Press.
- (2015) 'Autonomy in Moral and Political Philosophy' *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.) URL = <<https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/>>.
- Christman, John and Anderson, Joel, eds. (2005) *Autonomy and the Challenges to Liberalism. New Essays*. Cambridge University Press.
- Clarke, Randolph (1993) 'Toward A Credible Agent-Causal Account of Free Will' *Noûs*, 27:2, 191–203.
- (2003) *Libertarian Accounts of Free Will*. Oxford University Press.
- Clausen, Andrea (2004) *How can conceptual content be social and normative, and, at the same time, be objective?* Ontos Verlag.
- Clegg, Joshua W., ed. (2013) *Self-Observation in the Social Sciences*. Transaction Publishers.
- Clegg, Stewart R. and Haugaard, Mark, eds. (2009) *The SAGE Handbook of Power*. Sage.
- Code, Lorraine (1993) 'Taking Subjectivity into Account' in Alcoff and Potter (1993), Chapter 2, 15–48.
- Cohen, L. Jonathan (1992) *An Essay on Belief and Acceptance*. Clarendon Press.
- Colapietro, Vincent (2000) 'Further Consequences of a Singular Capacity' in Muller and Brent (2000), Chapter 9, 136–158.

- Colburn, Ben (2010) *Autonomy and Liberalism*. Routledge.
- Collins, Arthur W. (1987) *The Nature of Mental Things*. University of Notre Dame Press.
- Conly, Sarah (2013) *Against Autonomy. Justifying Coercive Paternalism*. Cambridge University Press.
- Connerton, Paul, ed. (1976) *Critical Sociology. Selected Readings*. Penguin Books.
- Coole, Diana (2013) 'From Rationalism to Micro-power. Freedom and Its Enemies' in Bruce Baum and Robert Nichols (2013), 199–215.
- Coons, Christian and Michael Weber, eds. (2014a) *Manipulation. Theory and Practice*. Oxford University Press.
- (2014b) 'Introduction. Manipulation. Investigating the core concept and its moral status' in Coons and Weber (2014a), 1–16.
- Cooper, Cary L. and Pervin, Lawrence A., eds. (1987) *Personality. Critical Concepts in Psychology. Volume III: Issues in Personality Theory and Research*. Routledge.
- Cooper, John M. (2003) 'Stoic Autonomy' in Paul, Miller and Paul (2003), 1–29.
- Cope, Zak (2008) *Dimensions of Prejudice. Towards a Political Economy of Bigotry*. Peter Lang.
- Costall, Alan (2006) 'Introspectionism' and the mythical origins of scientific psychology' *Consciousness and Cognition*, 15, 634–654.
- Cotta, Sergio (1983) 'Persona (filosofia del diritto)' in Mortati and Santoro-Passarelli (1983), 159–169.
- Counts, I.W.; W. Counts and Robert S. McCord (1999) *A Life Is More than a Moment: The Desegregation of Little Rock's Central High*.
- Cowie, Fiona (1999) *What's Within? Nativism Reconsidered*. Oxford University Press.
- Crane, Tim and Patterson, Sarah, eds. (2000) *History of the Mind-Body Problem*. Routledge.
- Crittenden, Jack (2002) *Democracy's Midwife. An Education in Deliberation*. Lexington Books.
- Crocker, Lawrence (1980) *Positive Liberty. An Essay in Normative Political Philosophy*. Martinus Nijhoff Publishers.
- Crombie, Alexander (1793/ 1989) *An essay on philosophical necessity*. Thoemmes.
- Crosby, Faye; Bromley, Stephanie and Saxe, Leonard (1980) 'Recent Unobtrusive Studies of Black and White Discrimination and Prejudice: A Literature Review' *Psychological Bulletin*, 87:3, 546–563.
- Crossley, Nick (2013a) 'Habitus and Habitus' *Body & Society*, 19:2-3, 136–161.
- (2013b) 'Pierre Bourdieu's *Habitus*' in Sparrow and Hutchinson (2013), Chapter 13, 291–307.
- Crowder, George (2004) *Isaiah Berlin: Liberty, Pluralism and Liberalism*. Polity Press.
- Crowder, George and Hary, Henry, ed. (2007) *The One and the Many: Reading Isaiah Berlin*. Prometheus Books.
- Cudd, Ann and Eftekhari, Seena (2018) 'Contractarianism' *The Stanford Encyclopedia of*

- Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/sum2018/entries/contractarianism/>>.
- Cullity, Garrett and Gaut, Berys, eds. (1997) *Ethics and Practical Reason*. Clarendon Press.
- Cuypers, Stefaan E. (1993) 'Searle, Dennett and Davidson on Original Intentionality' in Stelzner (1993), 215–225.
- (2000) 'Autonomy Beyond Voluntarism: In Defense of Hierarchy' *Canadian Journal of Philosophy*, 30:2, 225–256.
- Czarniawska, Barbara (2004) *Narratives in Social Science Research*. Sage Publications.
- Darwall, Stephen (1995) *The British Moralists and the Internal 'Ought': 1640-1740*. Cambridge: Cambridge University Press.
- (2006) *The Second Person Standpoint: Morality, Respect, and Accountability*. Cambridge.
- (2012) 'Grotius at the Creation of Modern Moral Philosophy' *Archiv für Geschichte der Philosophie*, 94(3), 296–325.
- Dasgupta, Nilanjana and Greenwald, Anthony G. (2001) 'On the Malleability of Automatic Attitudes: Combating Automatic Prejudice With Images of Admired and Disliked Individuals' *Journal of Personality and Social Psychology*, 81:5, 800–814.
- Daston, Lorraine (1978) 'British responses to Psycho-Physiology, 1860–1900' *Isis*, 69:2, 192–208.
- (1998) 'The Nature of Nature in Early Modern Europe' *Configurations*, 6:2, 149–172.
- Daston, Lorraine and Stolleis, Michael (2008a) 'Introduction: Nature, Law and Natural Law in Early Modern Europe' in Daston and Stolleis (2008b), 1–12.
- ,eds. (2008b) *Natural Law and Laws of Nature in Early Modern Europe. Jurisprudence, Theology, Moral and Natural Philosophy*. Ashgate.
- Davies, Catherine Glyn (2008) *Conscience as consciousness. The idea of self-awareness in French philosophical writing from Descartes to Diderot*. Voltaire Foundation.
- Davis, Joseph E. (2002a) *Stories of Change. Narrative and Social Movements*. State University of New York Press.
- ,ed. (2002b) 'Narrative and Social Movements. The Power of Stories' in Davis (2002a), 3–29.
- Davis, Winston, ed. (2001) *Taking Responsibility. Comparative Perspectives*. University Press of Virginia.
- De Caro, Mario and Voltolini, Alberto (2010) 'Is Liberal Naturalism Possible?' in De Caro and Macarthur (2010), 69 – 86.
- De Caro, Mario and Macarthur, David (2010) *Naturalism and Normativity*. Columbia University Press.
- De Dijn, Annelien (2020) *Freedom*. Harvard University Press.

- De Lauretis, Teresa (2000) 'Gender, Body, and Habit Change' in Muller and Brent (2000), Chapter 10, 159–175.
- Delgado, Richard (1989) 'Storytelling for Oppositionists and Others: A Plea for Narrative' *Michigan Law Review*, 87:8, 2411–2441.
- Dennett, Daniel C. (1971) 'Intentional Systems' *The Journal of Philosophy*, 68:4, 87–106.
- (1973) 'Mechanism and responsibility' in Honderich (1973), 159–184.
- (1976) 'Conditions of Personhood' in Dennett (1978) *Brainstorms. Philosophical Essays on Mind and Psychology*. Bradford Books, 267–285.
- (1980) 'The Milk of Human Intentionality' *The Behavioral and Brain Sciences*, 3, 428–430.
- (1984) 'I Could not have Done Otherwise – So What?' *The Journal of Philosophy*, 81:10, 553–565.
- (1984/ 2015) *Elbow Room. The Varieties of Free Will Worth Wanting*. New Edition. The MIT Press.
- (1987) *The Intentional Stance*. The MIT Press.
- (1990) 'The Interpretation of Texts, People and Other Artifacts' *Philosophy and Phenomenological Research*, 50, Supplement: 177–194.
- (1990) 'The Myth of Original Intentionality' in Mohyeldin Said, Newton-Smith, Viale and Wilkes (1990), 43–62.
- De Villiers, D.E. (2012) 'Die verantwoordelike-etiek van Max Weber: 'n Toepaslike etiek vir ons tyd?' *Nederduitse Gereformeerde Teologiese Tydskrif*, 53:3, 25–32.
- Dewey, John (1922) *Human Nature and Conduct* in Dewey (1983) *The Middle Works, 1899–1924. Volume 14: 1922*. Southern Illinois University Press.
- (1929) *The Quest for Certainty* in Dewey (1984) *The Later Works, 1925–1953. Volume 4: 1929*. Southern Illinois University Press
- Descombes, Vincent (1995) *La Denrée Mentale*. Les Éditions de Minuit.
- Devine, Patricia G. (1989) 'Stereotypes and Prejudice: Their Automatic and Controlled Components' *Journal of Personality and Social Psychology*, 65:1, 5–18.
- Devine, Patricia G.; Forscher, Patrick S.; Austin, Anthony J. and Cox, William T.L. (2012) 'Long-term reduction in implicit race bias: A prejudice habit-breaking intervention' *Journal of Experimental Social Psychology*, 48: 1267–1278.
- Devroye, Jennifer (2009) 'The Case of D.H. and Others v. the Czech Republic' *Northwestern Journal of International Human Rights*, 7:1, 81–101.
- Dieguez, Sebastian and Annoni, Jean-Marie (2013) 'Asomatognosia: disorders of the bodily self' in Godefroy (2013), Chapter 14, 170–192.
- Dijksterhuis, E.J. (1950) *De Mechanisering van het Wereldbeeld*. Meulenhoff.
- Dijksterhuis, Ap and Loran F. Nordgren (2006) 'A Theory of Unconscious Thought'. *Perspectives on Psychological Science*, 1:2, 95–109.
- Dixon, John and Levine, Mark, eds. (2012) *Beyond Prejudice. Extending the Social*

- Psychology of Conflict, Inequality and Social Change*. Cambridge University Press.
- Donert, Celia (2017) *The Rights of the Roma. The Struggle for Citizenship in Postwar Czechoslovakia*. Cambridge University Press.
- Doris, John M. (2015) *Talking to Our Selves. Reflection, Ignorance, and Agency*. Oxford University Press.
- Dorschel, Andreas (2001) *Nachdenken über Vorurteile*. Felix Meiner Verlag.
- Dovidio, John F. (2001) 'On the Nature of Contemporary Prejudice: The Third Wave' *Journal of Social Issues*, 57:4, 829–849.
- Dovidio, John F. and Gaertner, Samuel L. (2004) 'Aversive Racism' *Advances in Experimental Social Psychology*, 36, 1–52.
- Dovidio, John F.; Glick, Peter and Rudman, Laurie A. (2005a) 'Introduction: Reflecting on *The Nature of Prejudice: Fifty Years after Allport*' in Dovidio, Glick, Rudman and Laurie (2005b), 1–16.
- ,eds. (2005b) *On the Nature of Prejudice. Fifty Years after Allport*. Blackwell Publishing.
- Dowe, Phil (2002) 'What is Determinism?' in Atmanspracher and Bishop (2002), 309–319.
- Dreon, Roberta (2016) 'Understanding rules as habits. Developing a pragmatist anthropological approach' *Paradigmi*, XXIV:3, 103–117.
- Drewniak, Daniel; Krones, Tanja; Sauer, Carsten and Wild, Verina (2016) 'The influence of patients' immigration background and residence permit status on treatment decisions in health care. Results of a factorial survey among general practitioners in Switzerland' *Social Science & Medicine*, 161, 64–73.
- Dreyfus, Hubert L. (1991) *Being-in-the-World. A Commentary on Heidegger's Being and Time*. Cambridge, Massachusetts: The MIT Press.
- Du Bois, W.E.B. (1898) 'The Study of the Negro Problems' *The Annals of the American Academy of Political and Social Science*, 11, 1–23.
- (1903/ 2007) *The Souls of Black Folk*. Oxford University Press.
- (1920/ 2007) *Darkwater. Voices from Within the Veil*. Oxford University Press.
- Duckitt, J. (1992) *The social psychology of prejudice*.
- Dummett, Michael (1973) *Frege: Philosophy of Language*. Harper & Row.
- Durham, Michael S. (1991) *Powerful Days. The Civil Rights Photography of Charles Moore*. Stewart, Tabori & Chang.
- Durrheim, Kevin (2012) 'Implicit prejudice in mind and interaction' in Dixon and Levine (2012), 179–199.
- Dworkin, Gerald (1970) 'Acting Freely' *Noûs*, 4:4, 367–383.
- (1976) 'Autonomy and Behavior Control' *The Hastings Center Report*, 6:1, 23–28.
- (1981/ 1989) 'The Concept of Autonomy' in John Christman, ed. (1989) *The Inner Citadel. Essays on Individual Autonomy*, 54–62. Oxford University Press.
- (1988) *The Theory and Practice of Autonomy*. Cambridge University Press.

- Earman, John (1986) *A Primer on Determinism*. D. Reidel Publishing Company.
- (2004) ‘Determinism: What We Have Learned and What We Still Don’t Know’ in Campbell, O’Rourke and Shier (2004), Chapter 1, 21–46.
- Edwards, Jonathan (1754/ 1957) *Freedom of the Will. The Works of Jonathan Edwards, Volume 1*. Yale University Press.
- Eidelson, Benjamin (2013) ‘Treating People as Individuals’ in Hellman and Moreau (2013), Chapter 10, 203–227.
- Eilan, Naomi (2005) ‘Joint Attention, Communication, and Mind’ in Eilan; Hoerl; McCormack, and Roessler (2005), Chapter 1, 1–33.
- Eilan, Naomi; Hoerl, Christoph; McCormack, Teresa and Roessler, Johannes, eds. (2005) *Joint Attention: Communication and Other Minds. Issues in Philosophy and Psychology*. Clarendon Press.
- Ellenberger, Henri F. (1970) *The Discovery of the Unconscious. The History and Evolution of Dynamic Psychiatry*. BasicBooks.
- Ellis, Elisabeth, ed. (2012) *Kant’s Political Theory. Interpretations and Applications*. The Pennsylvania State University Press.
- Ellison, Ralph (1947/ 1994) *Invisible Man*. The Modern Library.
- (1989) ‘On being the target of discrimination’ in Ralph Ellison (1995), *The Collected Essays of Ralph Ellison*. The Modern Library.
- Engel, Pascal (2002) ‘Free Believers’ *Manuscrito*, 25:3, 155–175.
- Epley, Nicholas; Savitsky, Kenneth and Kacheliski, Robert A. (1999) ‘What Every Skeptic Should Know About Subliminal Persuasion’ in Nier (2013), 313–324.
- Erdelack, Wesley (2011) ‘Antivoluntarism and the birth of autonomy’ *Journal of Religious Ethics*, 39:4, 651–679.
- Erman, Eva and Möller, Niklas (2014) ‘Debate: Brandom and political philosophy’ *Journal of Political Philosophy*, 22:4, 486–498.
- (2015) ‘What not to expect from the pragmatic turn in political theory’ *European Journal of Political Theory*, 14:2, 121–140.
- Eskew, Glenn T. (1997) *But for Birmingham. The Local and National Movements in the Civil Rights Structure*. The University of North Carolina Press.
- Estlund, David, ed. (2012) *The Oxford Handbook of Political Philosophy*. Oxford University Press.
- European Commission (2012) *National Roma Integration Strategies: a first step in the implementation of the EU Framework*. COM(2012) 226 final.
- European Court of Human Rights (2007) *Case of D.H. and Others v. the Czech Republic*, 13 November 2007, application number 57325/00, Grand Chamber of the European Court of Human Rights. Available at <https://hudoc.echr.coe.int>.
- European Roma Rights Centre (2011) *Breaking the Silence. A Report by the European Roma Rights Centre and People in Need. Trafficking in Romani Communities*.
- European Roma Rights Centre, Amnesty International Czech Republic and Hate is

- No Solution Coalition (2012) ‘Discriminatory Violence in Czech Republic’; an open letter, dated 1 March 2012, to Mr. Jaromír Dvořák, Mr. Milan Boček, Mr. Martin Louka, Mr. Jaroslav Sykáček Mr. Josef Hýbl, Ms. Eva Džumanová, Ms. Alice Zemanová, Mr. Jan Kubice, Mr. Petr Lessy.
- European Union Agency for Fundamental Rights (2014) ‘Education: the situation of Roma in 11 Member States’ https://fra.europa.eu/sites/default/files/fra-2014_roma-survey_education_tk0113748enc.pdf (accessed on 4 September 2020).
- (2018) *Handbook on European non-discrimination law. 2018 edition*. Luxembourg: Publications Office of the European Union.
- Evans, Jonathan St. B.T. (2008) ‘Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition’ *The Annual Review of Psychology*, 59, 255–278.
- Evans, Jonathan St. B.T. and Frankish, Keith, eds. (2009) *In Two Minds. Dual Processes and Beyond*. Oxford University Press,
- Faden, Ruth R. and Tom L. Beauchamp, in collaboration with Nancy M.P. King (1986) *A History and Theory of Informed Consent*. Oxford University Press.
- Faucher, Luc (2016) ‘Revisionism and Moral Responsibility for Implicit Attitudes’ in Brownstein and Saul (2016), Chapter 1.5, 115–144.
- Fazio, Russell H.; Jackson, Joni R.; Dunton, Bridget C. and Williams, Carol J. (1995) *Journal of Personality and Social Psychology*, 69:6, 1013–1027.
- Feinberg, Joel (1962) ‘Problematic Responsibility in Law and Morals’ *The Philosophical Review*, 71:3, 340–351.
- (1986) *Harm to Self. The Moral Limits of the Criminal Law. Volume Three*. Oxford University Press.
- (1986/ 1989) ‘Autonomy’ in John Christman, ed. (1989) *The Inner Citadel. Essays on Individual Autonomy*, 27–53. Oxford University Press.
- Fekete, Liz (2012) *Peddars of Hate: the violent impact of the European far Right*. The Institute of Race Relations.
- Feldman, Richard (2000) ‘The Ethics of Belief’. *Philosophy and Phenomenological Research*, LX:3, 667–696.
- Festenstein, Matthew (1997) *Pragmatism & Political Theory*. Polity Press.
- Feyerabend, Paul K. and Maxwell, Grover, eds. (1966) *Mind, Matter, and Method. Essays in Philosophy and Science in Honor of HERBERT FEIGL*. University of Minnesota Press.
- Ffytche, Matt (2012) *The Foundation of the Unconscious. Schelling, Freud and the Birth of the Modern Psyche*. Cambridge University Press.
- Fischer, Alexander (2017) *Manipulation. Zur Theorie und Ethik einer Form der Beeinflussung*. Suhrkamp.
- Fischer, John Martin (1982) ‘Responsibility and Control’. *The Journal of Philosophy*, 79:1, 24–40.
- (1987) ‘Responsiveness and Moral Responsibility’ in Schoeman (1987), 81–106.

- (1999) ‘Recent Work on Moral Responsibility’ *Ethics*, 110:1, 93–139.
- (2005a) ‘General introduction’ in Fischer (2005b), xxiii–xxx.
- ed. (2005b) *Free Will. Critical Concepts in Philosophy. Volume I. Concepts and Challenges*. Routledge.
- (2006) *My Way. Essays on Moral Responsibility*. Oxford University Press.
- (2016) ‘How Do Manipulation Arguments Work?’ *The Journal of Ethics*, 20:1–3, 47–67.
- Fischer, John Martin and Mark Ravizza, eds. (1993) *Perspectives on Moral Responsibility*. Cornell University Press.
- (1998) *Responsibility and Control. A Theory of Moral Responsibility*. Cambridge University Press.
- Fisher, Walter R. (1987) *Human Communication As Narration: Toward a Philosophy of Reason, Value, and Action*. University of South Carolina Press.
- Flikschuh, Katrin (2000) *Kant and modern political philosophy*. Cambridge University Press.
- Finlayson, James Gordon (2005) ‘Habermas’s Moral Cognitivism and the Frege-Geach Challenge’ *European Journal of Philosophy*, 13:3, 319–45.
- Fiske, Susan T. (2000) ‘Stereotyping, prejudice, and discrimination at the seam between the centuries: evolution, culture, mind, and brain’. *European Journal of Social Psychology*, 30, 299–322.
- Fodor Jerry A. (1983) *The Modularity of Mind*. The MIT Press.
- (1998) *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.
- Foisneau, Luc, ed. (1997) *Politique, droit et théologie chez Bodin, Grotius et Hobbes*. Éditions Kimé.
- Foucault, Michel (1966) *Les mots et les choses*. Gallimard.
- Fossen, Thomas (2011) *Political Legitimacy and the Pragmatic Turn*. Dissertation.
- (2014) ‘Politicizing Brandom’s Pragmatism: Normativity and the Agonal Character of Social Practice’ *European Journal of Philosophy*, 22:3, 371–395.
- (2019) ‘Language and legitimacy: Is pragmatist political theory fallacious?’ *European Journal of Political Theory*, 18:2, 293–305.
- Foster, Kenneth R. (2006) ‘Engineering the brain’ in Illes (2006), Chapter 13, 185–199.
- Frank, Manfred (2015) *Präreflexives Selbstbewusstsein. Vier Vorlesungen*. Reclam.
- Frankenberg, Günter and Rödel, Ulrich (1981) *Von der Volkssouveränität zum Minderheitenschutz. Die Freiheit politischer Kommunikation im Verfassungsstaat*. Europäische Verlagsanstalt.
- Frankfurt, Harry G. (1969) ‘Alternate Possibilities and Moral Responsibility’ *The Journal of Philosophy*, 66:23, 829–839.
- (1971) ‘Freedom of the Will and the Concept of a Person’ *The Journal of Philosophy*, 68:1, 5–20.
- (1987) ‘Identification and Wholeheartedness’ in Schoeman (1987).

- Frankish, Keith (2010) 'Dual-Process and Dual-System Theories of Reasoning' *Philosophy Compass*, 5:10, 914–926.
- (2004) *Mind and Supermind*. Cambridge University Press.
- Frankish, Keith and Jonathan St.B.T. Evans (2009) 'The duality of mind: An historical perspective' in Evans and Frankish (2009), Chapter 1, 1–29.
- Frege, Gottlob (1879) 'Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens' in Frege (1977), V–88.
- (1977) *Begriffsschrift und andere Aufsätze*. Wissenschaftliche Buchgesellschaft.
- Fricker, Elizabeth (2006) 'Testimony and Epistemic Autonomy' in Jennifer Lackey and Ernest Sosa, eds. (2006), Chapter 10.
- Fricker, Miranda (2007) *Epistemic Injustice. Power and the Ethics of Knowing*. Oxford University Press.
- (2016a) 'Fault and No-Fault Responsibility for Implicit Prejudice. A Space for Epistemic "Agent-Regret"' in Brady and Fricker (2016), Chapter 2, 33–50.
- (2016b) 'Epistemic Injustice and the Preservation of Ignorance' in Peels and Blauw (2016), Chapter 9, 160–177.
- Fricker, Miranda; Peter J. Graham, David Henderson and Nikolaj J.L.L. Pedersen, eds. (2020) *The Routledge Handbook of Social Epistemology*. Routledge.
- Friedman, Marilyn A. (1986) 'Autonomy and the Split-Level Self' *The Southern Journal of Philosophy*, XXIV:1, 19–35.
- Fulkerson, Richard P. (1979) 'The public letter as a rhetorical form: structure, logic, and style in King's "Letter from Birmingham jail"' *The Quarterly Journal of Speech*, 65:2, 121–136.
- Gadamer, Hans Georg (1960/ 1990) *Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik*. Mohr Siebeck.
- Gaertner, Samuel L. and McLaughlin, John P. (1983) 'Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics' *Social Psychological Quarterly*, 46:1, 23–30.
- Galipeau, Claude J. (1994) *Isaiah Berlin's Liberalism*. Clarendon Press.
- Gallagher, Shaun and Zahavi, Dan (2012) *The Phenomenological Mind*. Second edition. Routledge.
- Gallate, J.; Wong, C.; Ellwood, S.; Chi, R. and Snyder, A. (2011) 'Noninvasive brain stimulation reduces prejudice scores on an implicit association test' *Neuropsychology*, 25, 185–92.
- Galston, William A. (1982) 'Moral Personality and Liberal Theory: John Rawls's "Dewey Lectures"' *Political Theory*, 10: 4, 492–519.
- Garrett Aaron, ed. (2014) *The Routledge Companion to Eighteenth Century Philosophy*. Routledge.
- Garvey, Stephen P. (2008) 'Self-Defense and the Mistaken Racist' *New Criminal Law Review*, 11:1, 119–171.

- (2015) ‘Injustice, Authority, and the Criminal Law’ in Sarat (2015), 42–81.
- Gasché, Rodolphe (1986) *The Tain of the Mirror. Derrida and the Philosophy of Reflection*. Harvard University Press.
- Gatens, Moira (1996) *Imaginary Bodies: Ethics, Power and Corporeality*. Routledge.
- Gauthier, David (1986) *Morals by Agreement*. Oxford University Press.
- Gawronski, Bertram (2009) ‘Ten Frequently Asked Questions About Implicit Measures and Their Frequently Supposed, But Not Entirely Correct Answers’ *Canadian Psychology/Psychologie canadienne*, 50:3, 141–150.
- Gawronski, Bertram and Creighton, Laura A. (2013) ‘Dual Process Theories’ in Donal E. Carlston, ed. (2013) *The Oxford Handbook on Social Cognition*, 282–312. Oxford University Press.
- Gawronski, Bertram; Hofmann, Wilhelm and Wilbur, Christopher J. (2006) ‘Are “implicit” attitudes unconscious?’ *Consciousness and Cognition*, 15, 485–499.
- Gawronski, Bertram and Payne, B. Keith, eds. (2010) *Handbook of Implicit Social Cognition. Measurement, Theory, and Applications*. The Guilford Press.
- Gawronski, Bertram; Sherman, Jeffrey W. and Trope, Yaacov (2014) ‘Two of What? A Conceptual Analysis of Dual-Process Theories’ in Sherman, Gawronski and Trope (2014), 3–19.
- Geach, P.T. (1960) ‘Ascriptivism’ in Geach (1972), 250–254.
- (1965) ‘Assertion’ in Geach (1972), 254–269.
- (1972) *Logic Matters*. University of California Press.
- Geertz, Clifford (1973) *The Interpretation of Cultures*. Basic Books.
- Gendler, Tamar Szabó (2008) ‘Alief and Belief’ *The Journal of Philosophy*, 105:10, 634–663.
- Gennaro, Rocco J., ed. (2004) *Higher-Order Theories of Consciousness. An Anthology*. John Benjamins Publishing Company.
- Gerrans, Philip (2000) ‘Refining the Explanation of Cotard’s Delusion’ *Mind & Language*, 15:1, 111–122.
- (2005) ‘Tacit knowledge, rule following and Pierre Bourdieu’s philosophy of social science’ *Anthropological Theory*, 5:1, 53–74.
- Gethmann, Carl Friedrich (1979) *Protologik. Untersuchungen zur formalen Pragmatik von Begründungsdiskursen*. Suhrkamp Verlag.
- (2007) *Vom Bewusstsein zum Handeln. Das phänomenologische Projekt und die Wende zur Sprache*. Wilhelm Fink Verlag.
- Gettier, Edmund L. (1963) ‘Is Justified True Belief Knowledge?’ *Analysis*, 23:6, 121–123.
- Geulincx, Arnold (1665) *Annotata ad Ethicam* in Geulincx (1968).
- (1968) *Sämtliche Schriften in fünf Bänden. Dritter Band*. Friedrich Fromman Verlag.
- (2006) *Ethics. With Samuel Beckett’s Notes*. Brill.

- Gibbons, Michael T., ed. (2015) *The Encyclopedia of Political Thought. First Edition*. John Wiley & Sons.
- Gigerenzer, Gerd; Zeno Swijtink; Theodore Porter; Lorraine Daston; John Beatty and Lorenz Krüger (1989) *The Empire of Chance. How probability changed science and everyday life*.
- Gilbert, Margaret (1989) *On Social Facts*. Princeton University Press.
- (2000) *Sociality and Responsibility. New Essays in Plural Subject Theory*. Rowman and Littlefield Publishers.
- Gill, Christopher (1988) 'Personhood and Personality: The Four-personae Theory in Cicero, de Officiis I' in *Oxford Studies in Ancient Philosophy, Volume VI*, 169–199.
- ed. (1990) *The Person and the Human Mind. Issues in Ancient and Modern Philosophy*. Clarendon Press.
- Gilman, Sander L. and Thomas, James M. (2016) *Are Racists Crazy? How Prejudice, Racism, and Antisemitism Became Markers of Insanity*. New York University Press.
- Giusberti, Franco (1982) *Materials for a Study on Twelfth Century Scholasticism*. Bibliopolis.
- Glasgow, Joshua (2016) 'Alienation and responsibility' in Brownstein and Saul (2016), Chapter 1.2, 37–61.
- Glüer, Kathrin (1999) *Sprache und Regeln. Zur Normativität von Bedeutung*. Akademie Verlag.
- (2002) 'Explizites und implizites Regelfolgen'.
- Glüer, Kathrin and Peter Pagin (1999) 'Rules of meaning and practical reasoning' *Synthese*, 117, 207–227.
- Glüer, Kathrin and Åsa Wikforss (2009) 'Against Content Normativity' *Mind*, 118, 31–70.
- Gödde, Günter (1999) *Traditionslinien des "Unbewußten". Schopenhauer – Nietzsche – Freud*. Edition diskord.
- Godefroy, Olivier, ed. (2013) *The Behavioral and Cognitive Neurology of Stroke*. Second Edition. Cambridge University Press.
- Godel, Rainer (2007) *Vorurteil – Anthropologie – Literatur. Der Vorurteilsdiskurs als Modus der Selbstaufklärung im 18. Jahrhundert*. Niemeyer.
- Göhler, Gerhard (2009) 'Power to' and 'Power over' in Clegg and Haugaard (2009), 27–39.
- Goldman, Loren (2015) 'Pragmatism' in Gibbons (2015).
- Goldston, James A. (2005) 'European Court to Address Racism in Landmark Cases', <https://www.justiceinitiative.org/voices/european-court-address-racism-landmark-cases>, 28 February (accessed on 31 August 2020).
- Goodin, Robert E. (1980) *Manipulatory Politics*. Yale University Press.
- Goodnow, Trischa (2003) 'Evaluating the Story. News Photographs and Social Narratives' *Visual Communication Quarterly*, 10:3, 4–9.

- (2005) ‘Using Narrative Theory to Understand the Power of News Photographs’ in Smith et al. (2005), 351–361.
- (2020) ‘Narrative Theory. Visual Storytelling’ in Josephson et al. (2020), 265–274.
- Goodwin, Morag (2006) ‘*D.H. and Others v. Czech Republic*: a major set-back for the development of non-discrimination norms in Europe’ *German Law Journal*, 7:4, 421–432.
- (2009) ‘Taking on racial segregation: the European Court of Human Rights at a *Brown v. Board of Education* moment?’ *Rechtsgeleerd Magazijn THEMIS*, 2009:3, 114–126.
- Gopnik, Alison (1993) ‘How we know our minds: The illusion of first-person knowledge of intentionality’ *Behavioral and Brain Sciences*, 16, 1–14.
- Gordley, James (1991) *The Philosophical Origins of Modern Contract Doctrine*. Clarendon Press.
- Gorin, Moti (2014) ‘Do Manipulators Always Threaten Rationality?’ *American Philosophical Quarterly*, 51:1, 51 – 61.
- Gould, Stephen Jay (1981/ 1996) *The Mismeasure of Man*. W.W. Norton & Company.
- Grau, Kurt Joachim (1916/ 1981) *Die Entwicklung des Bewusstseinsbegriffes im XVII. und XVIII. Jahrhundert*. Georg Olms Verlag.
- Green, T.H. (1886) ‘Lectures on the Principles of Political Obligation’ in Green (1986) *Lectures on the Principles of Political Obligation. And other writings*. Cambridge University Press.
- Greenberg, Jack (2004) *Brown v. Board of Education: Witness to a Landmark Decision. Anniversary Edition*. Twelve Tables Press.
- (2009–2010) ‘Roma Victimization: From Now to Antiquity’ *Columbia Human Rights Law Review*, 41:1, 1–12.
- (2010) ‘Report on Roma Education Today: From Slavery to Segregation and Beyond’ *Columbia Law Review*, 110:4, 919–1001.
- Greenspan, Patricia (2003) ‘The Problem With Manipulation’ *American Philosophical Quarterly*, 40:2, 155–164.
- Greenwald, Anthony G. (1992) ‘New look 3: Unconscious cognition reclaimed’ *American Psychologist*, 47:6, 766–779.
- Greenwald, Anthony G. and Banaji, Mahzarin R. (1995) ‘Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes’ *Psychological Review*, 102:1, 4–27.
- Greenwald, Anthony G. and Krieger, Linda Hamilton (2006) ‘Implicit Bias: Scientific Foundations’ *California Law Review*, 94:4, 945–967.
- Greenwald, Anthony G.; McGhee, Debbie E. and Schwartz, Jordan L.K. (1998) ‘Measuring Individual Differences in Implicit Cognition: The Implicit Association Test’ *Journal of Personality and Social Psychology*, 74:6, 1464–1480.
- Greenwald, Anthony G.; Poehlman, A. Andrew; Uhlmann, Eric Luis and Mahzarin, R. Banaji (2009) ‘Understanding and Using the Implicit Association Test: III. Meta-

- Analysis of Predictive Validity' *Journal of Personality and Social Psychology*, 97:1, 17–41.
- Grenfell, Michael, ed. (2012) *Pierre Bourdieu. Key Concepts*. Second Edition. Routledge.
- Greshake, Gisbert (1981) 'Die theologische Herkunft des Personbegriffs' in Pöltner (1981), 75–86.
- Grönert, Peter (2006) *Normativität, Intentionalität und praktische Schlüsse. Warum sich der menschliche Geist nicht reduzieren lässt*. Mentis Verlag.
- Gross, Matthias and McGoey, Linsey, eds. (2015) *Routledge International Handbook of Ignorance Studies*. Routledge.
- (1868) *De Jure Praedae. Commentarius. Ex Auctoris Codice descriptis et vulgavit*. Martinus Nijhoff.
- Gruber, David F. (1989) 'Foucault's Critique of the Liberal Individual' *The Journal of Philosophy*, 86:11, 615–621.
- Grundmann, Thomas; Hofmann, Frank; Misselhorn, Catrin; Waibel, Violetta L. and Zanetti, Véronique (2005) *Anatomie der Subjektivität*. Suhrkamp.
- Guelzo, Allen C. (1989) *Edwards on the Will. A Century of American Theological Debate*. Wipf & Stock.
- Gustafson, Donald A., ed. (1964) *Essays in Philosophical Psychology*. Macmillan.
- Haakonssen, Knud (2010) 'Natural Law and Personhood: Samuel Pufendorf on Social Explanation' *Max Weber Lecture No. 2010/06*. European University Institute.
- Habermas, Jürgen (1970) 'Systematically Distorted Communication' in Connerton (1976), 348–362.
- (2000) Review Article 'From Kant to Hegel: On Robert Brandom's Pragmatic Philosophy of Language' *European Journal of Philosophy*, 8:3, 322–355.
- (2005a) *Zwischen Naturalismus und Religion*. Suhrkamp.
- (2005b) 'Freiheit und Determinismus' in Habermas (2005), 155–186.
- (2007) 'The Language Game of Responsible Agency and the Problem of Free Will: How can epistemic dualism be reconciled with ontological monism?' *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action*, 10:1, 13–50.
- Hacking, Ian (1983) 'Nineteenth Century Cracks in the Concept of Determinism' *Journal of the History of Ideas*, 44:3, 455–475.
- Hahn, Lewis Edwin, ed. (1997) *The Philosophy of Roderick Chisholm*. Open Court.
- Hamilton, William, ed. (1872) *The Works of Thomas Reid, D.D., now fully collected, with selections from his unpublished letters*. MacLachlan and Stewart.
- Hammack, Philip L. (2018a) 'Social Psychology and Social Justice: Critical Principles and Perspectives for the Twenty-First Century' in Hammack (2018b), Chapter 1, 3–39.
- ed. (2018b) *The Oxford Handbook of Social Psychology and Social Justice*. Oxford University Press.

- Hammer, Espen, ed. (2007) *German Idealism. Contemporary Perspectives*. Routledge.
- Hammond, Corydon D. (2014) 'A Review of the History of Hypnosis Through the Late 19th Century' *American Journal of Clinical Hypnosis*, 56:2, 174–191.
- Hamowy, Ronald (1987) *The Scottish Enlightenment and the Theory of Spontaneous Order*. Southern Illinois University Press.
- Hampton, Jean (1986) *Hobbes and the Social Contract Tradition*. Cambridge University Press.
- Hanna, Robert (2008) 'Kant in the twentieth century' in Moran (2008), 149–203.
- Hardy, Henry, ed. (2002) *Liberty*. Oxford University Press.
- Harper, Krista (2012) 'Visual Interventions and the "Crises in Representation" in Environmental Anthropology: Researching Environmental Justice in a Hungarian Romani Neighborhood' *Human Organization*, 71:3, 292–305.
- Harris, James A. (2005) *Of Liberty and Necessity. The Free Will Debate in Eighteenth Century British Philosophy*. Oxford University Press.
- (2014) 'Liberty, necessity, and moral responsibility' in Garrett (2014), 320–337.
- Harrison, Peter (1995) 'Newtonian Science, Miracles, and the Laws of Nature' *Journal of the History of Ideas*, 56:4, 531–553.
- Hart, H.L.A. (1948 – 1949) 'The Ascription of Responsibility and Rights' Proceedings of the Aristotelian Society, New Series, 49, 171–194.
- (1968) *Punishment and Responsibility. Essays in the Philosophy of Law*. Clarendon Press.
- Hasker, William (1999) *The Emergent Self*.
- Hassin, Ran R.; James S. Uleman and John A. Bargh, eds. (2005) *The New Unconscious*. Oxford University Press.
- Hatfield, Gary (2011) 'Transparency of Mind: The Contributions of Descartes, Leibniz, and Berkeley to the Genesis of the Modern Subject' in Busche (2011), 361–375.
- Hattiangadi, Anandi (2003) 'Making It Implicit: Brandom on Rule Following'.
- (2006) 'Is Meaning Normative?' *Mind & Language*, 21:2, 220–240.
- (2007) *Oughts and Thoughts. Rule-Following and the Normativity of Content*.
- Haworth, Lawrence (1986) *Autonomy. An Essay in Philosophical Psychology and Ethics*. Yale University Press.
- Heal, Jane (1978) 'Common Knowledge' *The Philosophical Quarterly*, 28:111, 116–131.
- Heath, Joseph (2011) *Following the Rules. Practical Reasoning and Deontic Constraint*. Oxford University Press. Sara Heinämaa, Vili Lähteenmäki and Pauliina Remes, eds. (2007) *Consciousness. From Perception to Reflection in the History of Philosophy*, 267–285.
- Heinämaa, Sara; Lähteenmäki, Vili and Remes, Pauliina, eds. (2007) *Consciousness. From Perception to Reflection in the History of Philosophy*.
- Hellman, Deborah and Sophia Moreau, eds. (2013) *Philosophical Foundations of Discrimination Law*. Oxford University Press.

- Henriot, Jacques (1977) 'Note sur la date et le sens du mot "responsabilité"' Archives de Philosophie du Droit, 22, 59–62.
- Henriques, Julian (1984) 'Social psychology and the politics of racism' in Henriques et al. (1998), Chapter 2, 60–89.
- Henriques, Julian; Hollway, Wendy; Urwin, Cathy; Venn, Couze and Walkerdine, Valerie (1998) *Changing the Subject. Psychology, social regulation and subjectivity*. Routledge.
- Herbert, Gary B. (1996) 'John Locke: Natural Rights and Natural Duties' in *Jahrbuch für Recht und Ethik / Annual Review of Law and Ethics*, 4, Themenschwerpunkt: *Bioethik und Medizinrecht / Bioethics and the Law*, 591–613.
- Herrmann, Martina (2001) 'Der Personbegriff in der analytischen Philosophie' in Sturma (2001), 167–185.
- Hendricks, Gerret; Derick op de Graeff; Francis Daniell Pastorius and Abraham op den Graef (1898) 'Resolution of Germantown Mennonites' in Basker (2012), 1–3.
- Heyes, Cressida J., ed. (2003) *The Grammar of Politics. Wittgenstein and Political Philosophy*. Cornell University Press.
- Higgins, E. Tory and John A. Bargh (1992) 'Unconscious Sources of Subjectivity and Suffering: Is Consciousness the Solution?' in Martin and Tesser (1992), Chapter 3, 67–103.
- Hill, T.E. (1987) 'The Importance of Autonomy' in Kittay and Meyers (1987), Chapter 7, 129–138.
- (1989) 'The Kantian Conception of Autonomy' in Christman (1989), Chapter 6, 91–105.
- (1991) *Autonomy and self-respect*. Cambridge University Press.
- Hinchman, Lewis P. (1990) 'The Idea of Individuality: Origins, Meaning, and Political Significance' *The Journal of Politics*, 52:3, 759–781.
- (1996) 'Autonomy, Individuality, and Self-Determination' in James Schmidt, ed. (1996) *What Is Enlightenment? Eighteenth-Century Answers and Twentieth-Century Questions*, 488–516. University of California Press.
- Hirst, Paul and Bader, Veit, eds. (2001) *Associative Democracy: The Real Third Way*. Frank Cass Publishers.
- Hobbes, Thomas (1651/ 2012a) *Leviathan. Volume 2. The English and Latin Texts (i). The Clarendon Edition of the Works of Thomas Hobbes*. Clarendon Press.
- (1651/ 2012b) *Leviathan. Volume 3. The English and Latin Texts (ii). The Clarendon Edition of the Works of Thomas Hobbes*. Clarendon Press.
- Hodgkiss, Philip (2001) *The Making of the Modern World. The Surfacing of Consciousness in Social Thought*. The Athlone Press.
- Hoffmann, Thomas S. (2019). 'Spontaneität' in Ritter and Gründer (1971–2007), Volume 9, 1424–1434.
- Hoffmann, Tobias (2014) 'Intellectualism and voluntarism' in Pasnau and van Dyke

- (2014) Chapter 30, 414–427.
- Hofmann, Wilhelm and Timothy D. Wilson (2010) ‘Consciousness, Introspection, and the Adaptive Unconscious’ in Bertram Gawronski and B. Keith Payne, eds., *Handbook of Implicit Social Cognition. Measurement, Theory, and Applications*, 1–15. The Guilford Press.
- Holroyd, Jules (2011) ‘The Metaphysics of Relational Autonomy’ in Witt (2011), 99–115.
- (2012) ‘Responsibility for Implicit Bias’ *Journal of Social Philosophy*, 43:3, 274–306.
- (2017) ‘The social psychology of discrimination’ in Lippert-Rasmussen, Kasper, (ed.), Chapter 32.
- Holroyd, Jules and Katherine Puddifoot (2020) ‘Implicit Bias and Prejudice’ in Fricker et al. (2020), Chapter 31, 313–326.
- Holroyd, Jules; Robin Scaife and Tom Stafford (2017) ‘What is implicit bias?’ *Philosophy Compass*, 12:3, 1–13.
- Holthoorn, Frits van; Olson, David R., eds. (1987) *Common Sense. The Foundations for Social Science*. University Press of America.
- Holton, Robert (1997) ‘Bourdieu and Common Sense’ *SubStance*, 26:3, Issue 84, 38–52.
- Honderich, Ted, ed. (1973) *Essays on freedom of action*. Routledge & Kegan Paul.
- Honneth, Axel und Hans Joas, eds. (1986/ 2002) *Kommunikatives Handeln. Beiträge zu Jürgen Habermas’ “Theorie des kommunikativen Handelns”*. Erweiterte und aktualisierte Ausgabe. Suhrkamp.
- Hook, Sidney, eds. (1961) *Determinism and Freedom. In the the Age of Modern Science*. Collier Books.
- Hopman, Marieke J. (2016) *(W)elk kind heeft recht op onderwijs? Een onderzoek naar de betekenis van recht op onderwijs voor kinderen in Nederland, specifiek gericht op thuisonderwijs, thuiszitters en Roma kinderen*. Defence for Children International Nederland.
- Hospers, John (1958) ‘What Means This Freedom?’ in Hook (1961), 126 – 142.
- Hruschka, Joachim (1976) *Strukturen der Zurechnung*. Walter de Gruyter.
- (1986) ‘Imputation’ *Brigham Young University Law Review*, 669–710.
- Hubbard, Howard (1968) ‘Five long hot summers and how they grew’ *Public Interest*, 12, 3–24.
- Hume, David (1739/ 1978) *A Treatise of Human Nature*. Second edition. Oxford University Press.
- (1772-1777a/ 1994) ‘Of the Original Contract’ in Hume (1772-1777/ 1994b)
- (1772-1777b/ 1994) *Essays. Moral, Political, and Literary*. Liberty Fund.
- Hundert, E.J. (1987) ‘Enlightenment and the Decay of Common Sense’ in Holthoorn and Olson (1987), Chapter 8, 133–154.

- Hunter, Ian (2001) *Rival Enlightenments. Civil and Metaphysical Philosophy in Early Modern Germany*. Cambridge University Press.
- Hurlburt, Russell T. (1990) *Sampling Normal and Schizophrenic Inner Experience*. Plenum Press.
- Hurley, S.L. (2003) *Justice, Luck, and Knowledge*. Harvard University Press.
- Hutchison, Katrina and Jenkins, Fiona, eds. (2013) *Women in Philosophy. What Needs to Change?* Oxford University Press.
- Hydén, Margareta (1994) *Woman Battering as Marital Act. The Construction of a Violent Marriage*. Scandinavian University Press.
- Hyvärinen, Matti (2006) 'Towards a conceptual history of narrative' in M. Hyvärinen; A. Korhonen and J. Mykkänen (eds.) (2006) *Volume 1: The Travelling Concept of Narrative*, 20 – 41. Helsinki Collegium for Advanced Studies, University of Helsinki. <https://helda.helsinki.fi/handle/10138/25730> (accessed on 14 October 2020).
- Illes, Judy, ed. (2006) *Neuroethics. Defining the issues in theory, practice, and policy*. Oxford University Press.
- Irwin, T.H. (1980) 'Reason and Responsibility in Aristotle' in Rorty (1980), Chapter 8, 117–155.
- Insin, Engin F. (2008) 'Theorizing Acts of Citizenship' in Engin and Nielsen (2008), Chapter 1, 15–43.
- Insin, Engin F. and Greg M. Nielsen, eds. (2008) *Acts of Citizenship*. Zed Books.
- Jackson, Nicholas (2007) *Hobbes, Bramhall and the politics of Liberty and Necessity. A Quarrel of the Civil Wars and Interregnum*. Cambridge University Press.
- Jacoway, Elizabeth (2007) *Turn Away Thy Son. Little Rock, the crisis that shocked the nation*. Free Press.
- Jäger, Mechtild (2001) "Was ist Begründung?" Schwierigkeiten durch den Primat der Pragmatik vor der Semantik am Beispiel des Erlanger/Konstanzer Konstruktivismus' *Journal for General Philosophy of Science*, 32, 167–192.
- James, William (1884) 'The Dilemma of Determinism' in William James (1992) *Writings 1878–1899*, 566–594. Library of America.
- Jay, Mike (2012) *The Influencing Machine. James Tilly Matthews and the Air Loom*. Strange Attractor Press. (Previously published in 2003 as *The Air Loom Gang*.)
- Jenkins, Richard (1982) 'Critical Note. Pierre Bourdieu and the Reproduction of Determinism' *Sociology*, 16, 270 – 281.
- Joas, Hans (1986) 'Die unglückliche Ehe von Hermeneutik und Funktionalismus' in Honneth und Joas (1986/ 2002), 144–176.
- Johnson, Davi (2007) 'Martin Luther King Jr.'s 1963 Birmingham Campaign as Image Event' *Rhetoric & Public Affairs*, 10:1, 1–25.
- Johnson, Gabrielle M. (2020) 'The Psychology of Bias: From Data to Theory' in Beeghly and Madva (2020), 20–40.

- Johnston, David (1994) *The Idea of a Liberal Theory. A Critique and Reconstruction*. Princeton University Press.
- Jolls, Christine and Cass R. Sunstein (2006) 'The Law of Implicit Bias' *California Law Review*, 94, 969–996.
- Jones, Malcolm V. and Terry, Garth M., eds. (1983) *New Essays on Dostoyevsky*. Cambridge University Press.
- Jonsen, Albert R. (1968) *Responsibility in Modern Ethics*. Corpus Books.
- Josephson, Sheree; James D. Kelly and Ken Smith, eds. (2020) *Handbook of Visual Communication. Theory, Methods, and Media*. Routledge.
- Jost, John T. (2018) 'The IAT Is Dead, Long Live the IAT: Context-Sensitive Measures of Implicit Attitudes Are Indispensable to Social and Political Psychology' *Current Directions in Psychological Science*, 1–10.
- Jost, John T. and Banji, Mahzarin R. (1994) 'The Role of Stereotyping in System-Justification and the Production of false consciousness' *British Journal of Social Psychology*, 33, 1–27.
- Jost, John T.; Rudman, Laurie A.; Blair, Irene V.; Carney, Dana R.; Dasgupta, Nilanjana; Glaser, Jack and Hardin, Curtis D. (2009) 'The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore' *Research in Organizational Behavior*, 29, 39–69.
- Kagel, Martin (2010) 'A Blank Slate: Peregrination and Prejudice in Eighteenth-Century Travel Writing' *Publications of the English Goethe Society*, LXXIX:2, 79–94.
- Kahn, Jonathan (2017) 'Pills for Prejudice: Implicit Bias and Technical Fix for Racism' *American Journal of Law & Medicine*, 43 (2017): 263–278.
- (2018) *Race on the Brain. What Implicit Bias Gets Wrong About the Struggle for Racial Justice*. Columbia University Press.
- Kane, Robert (1996) *The Significance of Free Will*. Oxford University Press.
- Kant, Immanuel (1785/ 1786) *Grundlegung zur Metaphysik der Sitten* in Kant (1956), 7–102.
- (1797) *Die Metaphysik der Sitten* in Kant (1956), 303–634.
- (1784a) 'Beantwortung der Frage: Was ist Aufklärung?' in Kant (1964), 51–61.
- (1793) 'Über den Gemeinspruch: Das mag in der Theorie richtig sein, taugt aber nicht für die Praxis' in Kant (1964), 125–172.
- (1784b) 'Idee zu einer allgemeinen Geschichte in weltbürgerlicher Absicht' in Kant (1964), 33–50.
- (1956) *Werke in Sechs Bänden. Band IV. Schriften zur Ethik und Religionsphilosophie*. Wissenschaftliche Buchgesellschaft Darmstadt.
- (1964) *Werke in Sechs Bänden. Band VI. Schriften zur Anthropologie, Geschichtsphilosophie, Politik und Pädagogik*. Wissenschaftliche Buchgesellschaft Darmstadt.

- Karvayová, Magdaléna; Jolana Šmarhovyčová and Miroslav Klempár (2016) *Survey on the needs of Roma children enrolled to good quality primary schools in Ostrava since January 2014 until January 2015 and on identifying the difference in quality of education among the basic primary schools and segregated primary schools*. Available at https://lawenamenca.cz/_files/200000255-4159242537/Survey_Ostrava%202016.pdf (accessed on 1 April 2021).
- Kasher, Steven (1996) *The Civil Rights Movement. A photographic history, 1954–68*. Abbeville Press.
- Kaye, Sharon M. (2004) ‘Why the Liberty of Indifference Is Worth Wanting: Buridan’s Ass, Friendship, and Peter John Olivi’ *History of Philosophy Quarterly*, 21:1, 21 – 42.
- Kelly, Daniel and Erica Roedder (2008) ‘Racial Cognition and the Ethics of Implicit Bias’ *Philosophy Compass*, 3:3, 522–540.
- Kelly, Daniel; Luc Faucher and Edouard Machery (2010) ‘Getting Rid of Racism: Assessing Three Proposals in Light of Psychological Evidence’ *Journal of Social Philosophy*, 41:3, 293–322.
- Kelly, Robin D.G. (1994) *Race Rebels. Culture, Politics, and the Black Working Class*. The Free Press.
- Kenny, Anthony (1973) ‘Freedom, spontaneity and indifference’ in Honderich (1973), 89–104.
- Kent, Bonnie (1995) *Virtues of the Will. The Transformations of Ethics in the Late Thirteenth Century*. The Catholic University of America Press.
- Kernan, Ryan (2010) ‘Story/Discourse’ in Peter Melville Logan (2010) *The Encyclopedia of the Novel*. John Wiley & Sons, Ltd.
- Kershner, Stephen (2015) ‘Moral Responsibility and Foundationalism’ *Philosophia*, 43:2, 381–402.
- Keyssar, Alexander (2009) *The Right to Vote. The Contested History of Democracy in the United States*. Revised Edition. Basic Books.
- Kibble, Rodger (2014) ‘Discourse as practice: from Bourdieu to Brandom’; a paper (unpublished) presented at the *Proceedings of the 50th Anniversary Convention of the AISB* for the symposium *Questions, discourse and dialogue: 20 years after Making it Explicit*. See: <http://doc.gold.ac.uk/aisb50>.
- Kidd, Ian James; José Medina and Gaile Pohlhaus, Jr., eds. (2017) *The Routledge Handbook of Epistemic Injustice*. Routledge.
- Kihlstrom, John F. (1984) ‘Conscious, subconscious, unconscious: A cognitive perspective’ in Bowers and Meichenbaum (1984).
- (1987) ‘The Cognitive Unconscious’ in Cooper and Pervin (1987), 441–459.
- (1990) ‘The Psychological Unconscious’ in Pervin and John (1999), Chapter 17, 424–442.
- (2008) ‘The Automaticity Juggernaut – or: Are We Automatons After All?’ in Baer,

- Kaufman and Baumeister (2008), Chapter 8, 155–180.
- Kihlstrom, John F.; Terrence M. Barnhardt and Douglas J. Tataryn (1992) ‘The psychological unconscious: Found, lost, and regained’ *American Psychologist*, 47:6, 788 – 791.
- King, Anthony (2000) ‘Thinking with Bourdieu Against Bourdieu: A ‘Practical’ Critique of the Habitus’ *Sociological Theory*, 18:3, 417 – 433.
- King, Joyce E. (1991) ‘Dysconscious Racism: Ideology, Identity, and the Miseducation of Teachers’ *The Journal of Negro Education*, 60:2, 133–146.
- (1992) ‘Diaspora Literacy and Consciousness in the Struggle Against Miseducation in the Black Community’ *The Journal of Negro Education*, 61:3, Africentrism and Multiculturalism: Conflict or Consonance, 317–340.
- King, Martin Luther, Jr. (1963/ 2000) ‘Letter from Birmingham Jail’ in id. (1963/ 2000) *Why We Can't Wait*. New York: Penguin Random House LLC.
- King, Matt and Peter Carruthers (2012) ‘Moral Responsibility and Consciousness’ *Journal of Moral Philosophy*, 9, 200–228.
- King, Peter (2007) ‘Why isn't the mind-body problem Mediaeval?’ in Lagerlund (2007), 187–205.
- Kitchener, Richard F. (1994) ‘Semantic Naturalism: The Problem of Meaning and Naturalistic Psychology’ in Overton and Palermo (1994), Chapter 13, 279–293.
- Kitcher, Philip (1992) ‘The Naturalists Return’ *The Philosophical Review*, 101:1, 53–114.
- Kittay, Eva Feder and Meyers, Diana T. (1987) *Women and Moral Theory*. Rowman & Littlefield.
- Klein, D.B. (1977) *The Unconscious: Invention or Discovery? A Historico-Critical Inquiry*. Goodyear Publishing Company.
- Knapp, Mark L.; Hall, Judith A. and Horgan, Terrence G., eds. (2014) *Nonverbal Communication in Human Interaction*. Eighth Edition. Wadsworth Cengage Learning.
- Knesebeck, Julia von dem (2011) *The Roma Struggle for Compensation in Post-War Germany*. University of Hertfordshire Press.
- Knight, Carl and Stemplowska, Zofia (2011a) ‘Responsibility and Distributive Justice: An Introduction’ in Knight and Stemplowska (2011b), 1–23.
- eds. (2011b) *Responsibility and Distributive Justice*. Oxford University Press.
- Kobusch, Theo (1993) *Die Entdeckung der Person. Metaphysik der Freiheit und Modernes Menschenbild*. Wissenschaftliche Buchgesellschaft.
- (2006) ‘Person und Freiheit. Von der Rezeption einer vergessenen Tradition’ *Zeitschrift für Evangelische Ethik*, 50, 7–20.
- Kolb, Felix (2007) *Protest and Opportunities. The Political Outcomes of Social Movements*. Campus Verlag.
- Kontler, László and Somos, Mark, eds. (2018) *Trust and Happiness in the History of European Political Thought*. Brill.

- Kornblith, Hilary (2012) *On Reflection*. Oxford University Press.
- Korsgaard, Christine M. (1996a) *The Sources of Normativity*. Cambridge University Press.
- (1996b) *Creating the Kingdom of Ends*. Cambridge University Press.
- (1996b) ‘Morality as Freedom’ in Korsgaard (1996b), 159–187.
- Kovel, Joel (1970) *White Racism: A Psychohistory*.
- Kluger, Richard (1975/ 2004) *Simple Justice. The History of Brown v. Board of Education and Black America’s Struggle for Equality*. Vintage Books.
- Kriegel, Uriah (2012) ‘Moral Motivation, Moral Phenomenology, and the Alief/ Belief Distinction’ *Australasian Journal of Philosophy*, 90:3, 469–486.
- Kriegel, Uriah and Williford, Kenneth, eds. (2006) *Self-Representational Approaches to Consciousness*. The MIT Press.
- Kristjánsson, Kristján (1996) *Social Freedom. The responsibility view*. Cambridge University Press.
- Krizek, Robert L. (2017) ‘Narrative and Storytelling’ in Craig R. Scott and Laurie Lewis (Editors-in-Chief), James R. Barker, Joann Keyton, Timothy Kuhn, and Paaige K. Turner (Associate Editors) (2017) *The International Encyclopedia of Organizational Communication*. John Wiley & Sons.
- Kushner, Howard I. (1999) *A Cursing Brain? The Histories of Tourette Syndrome*. Harvard University Press.
- Kymlicka, Will (2002) *Contemporary Political Philosophy. An Introduction*. Second edition. Oxford University Press.
- Lackey, Jennifer and Ernest Sosa, eds. (2006) *The Epistemology of Testimony*. Oxford: Clarendon Press.
- Lacroix, Mariève (2009) ‘Les fondations épistémologiques de la responsabilité civile’ *Les Cahiers de droit*, 50:2, 415–433.
- LaFollette, Hugh, ed. (2013) *The International Encyclopedia of Ethics*. Blackwell Publishing Ltd.
- Lagerlund, H. (2007) *Forming the Mind: Essays on the Internal Senses and the Mind/Body Problem from Avicenna to the Medical Enlightenment*. Springer.
- Lai, Calvin K.; Hoffman, Kelly M. and Nosek, Brian A. (2013) ‘Reducing Implicit Prejudice’ *Social and Personality Psychology Compass*, 7:5, 315–330.
- Lance, Mark and W. Heath White (2007) ‘Stereoscopic vision: Persons, Freedom, and Two Spaces of Material Inference’ *Philosophers’ Imprint*, 7:4, 1–21.
- Lane, Kristin A.; Banaji, Mahzarin R.; Nosek, Brian A. and Greenwald, Anthony G. (2007) ‘Understanding and Using the Implicit Association Test: IV. What We Know (So Far) about the Method’ in Wittenbrink and Schwartz (2007), Chapter 3, 59–102.
- Lanting, K.; E. Marek; M. Ridder; Z. Feiszt; P. Verdonk; J. Suurmond (2018) ‘Amsterdam Public Health, Global Health program – health professionals’ perception of implicit

- bias towards minority patients in health care in the Netherlands and Hungary' *European Journal of Public Health*, 28, Supplement 1, 69–70.
- Larmore, Charles (1987) *Patterns of Moral Complexity*. Cambridge University Press.
- (2008) *The Autonomy of Morality*. Cambridge University Press.
- (2013) 'Kant and the Meanings of Autonomy' in Rush and Stolzenberg (2013), 3–21.
- Lattal, Kennon A. and Laipple, Joseph S. (2003) 'Pragmatism and Behavior Analysis' *Behavior Theory and Philosophy*, 41–61.
- LaVaque-Manty, Mika (2006) 'Kant's Children' *Social Theory and Practice*, 32:3, 365–388.
- Lavin, Chad (2008) *The Politics of Responsibility*. The University of Illinois Press.
- Lawrence III, Charles R. (1987) 'The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism' *Stanford Law Review*, 23:2, 317–388.
- (2008) 'Unconscious Racism Revisited: Reflections on the Impact and Origins of "The Id, the Ego, and Equal Protection"' *Connecticut Law Review*, 40:4, 931–978.
- (2015) 'Local kine implicit bias: Unconscious racism revisited (yet again)' *University of Hawai'i Law Review*, 37(2), 457–500.
- Leach, Colin Wayne (2002) 'II. The Social Psychology of Racism Reconsidered' *Feminism & Psychology*, 12:4, 439–444.
- (2005) 'Against the Notion of a "New Racism"' *Journal of Community & Applied Social Psychology*, 15, 432–445.
- Leahey, Thomas H. (1992) 'The Mythical Revolutions of American Psychology' *American Psychologist*, 47:2, 308–318.
- Leboeuf, Céline (2000) 'The Embodied Biased Mind' in Beeghly and Madva (2000), Chapter 2, 41–56.
- Legg, Catherine and Hookway, Christopher (2019) 'Pragmatism' *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/spr2019/entries/pragmatism/>>.
- Lehrer, Keith (2003) 'Reason and Autonomy' in Paul, Miller and Paul (2003), 177–198.
- Leschziner, Vanina (2019) 'Dual-process models in sociology' in Brekhuis, Wayne H. and Gabe Ignatow (2019), Chapter 10, 169–191.
- Levine, Michael P. and Pataki, Tamas, eds. (2004) *Racism in Mind*. Cornell University Press.
- Levy, Neil (2012) 'Implicit Bias and Moral Responsibility: Probing the Data'. Unpublished manuscript.
- (2014) *Consciousness and Moral Responsibility*. Oxford University Press.
- Levy, Neil (2015) 'Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements' *Noûs*, 49:4, 800–823.
- Levy, Neil and Michael McKenna (2009) 'Recent Work on Free Will and Moral Responsibility' *Philosophy Compass*, 4:1, 96–133.

- Lewis, David (1969) *Convention: A Philosophical Study*. Blackwell Publishers.
- (1979) ‘Scorekeeping in a Language Game’ *Journal of Philosophical Logic*, 8, 339–359.
- Lewis, Michael (1990) ‘The Development of Intentionality and the Role of Consciousness’. *Psychological Inquiry*, 1:3, 231–247.
- Liamputtong, Pranee (2007) *Researching the Vulnerable. A Guide to Sensitive Research Methods*. Sage Publications Ltd.
- Lichterman, Paul and Eliasoph (2014) ‘Civic Action’ *American Journal of Sociology*, 120:3, 798–863.
- Lindley, Richard (1986) *Autonomy*. MacMillan Education Ltd.
- Lippert-Rasmussen, Kasper (2015) ‘Discrimination: An Intriguing but Underexplored Issue in Ethics and Political Philosophy’ *Moral Philosophy and Politics*, 2:2, 207–217.
- ed. (2017) *Routledge Handbook on the Ethics of Discrimination*. London and New York: Routledge.
- Liston, Michael (2010) ‘Rabbits Astray and Significance Awandering: Review Essay on Mark Wilson’s Wandering Significance’ *Philosophy and Phenomenological Research*, 81:3, 809–817.
- Livingston, Alexander (2015) ‘Habit’ in Gibbons (2015).
- Locke, John (1690/ 1975). *An Essay concerning Human Understanding*. Clarendon Press.
- Loeffler, Ronald (2018) *Brandom*. Polity.
- López, Ian F. Haney (2000) ‘Institutional Racism: Judicial Conduct and a New Theory of Racial Discrimination’ *The Yale Law Journal*, 8, 1717–1884.
- Lucente, Gregory L. (1982) ‘Vico’s Notion of “Divine Providence” and the Limits of Human Knowledge, Freedom, and Will’ *MLN*, 97:1, Italian Issue, 183–191.
- Lueke, Adam and Gibson, Bryan (2016) ‘Brief Mindfulness Meditation Reduces Discrimination’ *Psychology of Consciousness: Theory, Research, and Practice*, 3:1, 34–44.
- Lukes, Steven (1974/ 2005) *Power. A Radical View*. Second Edition. Palgrave Macmillan.
- Lulé, D.; C. Zickler; S. Häcker; M.A. Bruno; A. Demertzi; F. Pellas; S. Laureys and A. Kübler (2009) ‘Life can be worth living in locked-in syndrome’. *Progress in Brain Research*, 177, 339–351.
- Lyons, John O. (1978) *The Invention of the Self. The Hinge of Consciousness in the Eighteenth Century*. Southern Illinois University Press.
- Lyons, William (1986) *The Disappearance of Introspection*. The MIT Press.
- MacCann, C. and Roberts, R.D. (2013) ‘Just as smart but not as successful: obese students obtain lower school grades but equivalent test scores to nonobese students’ *International Journal of Obesity*, 37, 40–46.
- Macdonald, Cynthia (2008) ‘Consciousness, Self-Consciousness, and Authoritative Self-Knowledge’ *Proceedings of the Aristotelian Society*, CVIII:3, 319–346.

- Machery, Edouard; Luc Faucher and Daniel R. Kelly (2010) 'On the Alleged Inadequacies of Psychological Explanations of Racism' *The Monist*, 93:2, 228–254.
- MacIntyre, Alasdair (1981/ 1985) *After Virtue*. Second Edition. Duckworth.
- (1982) 'How Moral Agents Became Ghosts or Why the History of Ethics Diverged from That of the Philosophy of Mind' *Synthese*, 53:2, 295–312.
- (1988) *Whose Justice? Which Rationality?* University of Notre Dame Press.
- Mackenzie, Catriona and Stoljar, Natalie, eds. (2000) *Relational Autonomy. Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford University Press.
- Macnamara, Coleen (2011) 'Holding others responsible' *Philosophical Studies*, 152, 81–102.
- MacMullan, Terrance (2009) *Habits of Whiteness. A Pragmatist Reconstruction*. Indiana University Press.
- Madva, Alex (2012) *The Hidden Mechanisms of Prejudice: Implicit Bias & Interpersonal Fluency*. Dissertation (unpublished).
- Mandelbaum, Eric (2016) 'Attitude, Inference, Association: On the Propositional Structure of Implicit Bias' *Noûs*, 50:3, 629–658.
- Manson, Neil Campbell (2002) "A tumbling-ground for whimsies"? The history and contemporary role of the conscious/unconscious contrast' in Crane and Patterson (2000), Chapter 6, 148–168.
- Marçal, Antonio Cota and Guilherme F.R. Kisteumacher (2010) 'Nichtpropositionalität and Propositionalität: Alternative oder komplementäre Formen des diskursiven Denkens?' in Bromand and Kreis (2010), 101–120.
- Marcoulatos, Iordanis (2003) 'John Searle and Pierre Bourdieu: Divergent Perspectives on Intentionality and Social Ontology' *Human Studies*, 26, 67–96.
- Marek, Erika; Timea Nemeth and Zsuzsa Orsos (2020) 'Implicit bias against the Romas in Hungarian healthcare: taboos or unrevealed areas for health promotion?' *Health Promotion International*, 1–9.
- Martin, Leonard L., Abraham Tesser, Abraham (1992) *The Construction of Social Judgments*. Lawrence Erlbaum Associates.
- Marvasti, Amir and McKinney, Karyn (2007) 'The Work of Making Racism Invisible' in Vera and Feagin (2007), Chapter 5, 67–78.
- Maton, Karl (2012) 'Habitus' in Grenfell (2012), Chapter 3, 48–64.
- Matson, Wallace (1966) 'Why isn't the mind-body problem ancient?' in Feyerabend and Maxwell (1996), 92–102.
- May, Larry and Stacey Hoffman, eds. (1991) *Collective Responsibility. Five Decades of Debate in Theoretical and Applied Ethics*. Rowman and Littlefield Publishers, Inc.
- McAdam, Doug (1983) 'Tactical Innovation and the Pace of Insurgency' *American Sociological Review*, 48:6, 735–754.
- (1996) 'The framing function of movement tactics: Strategic dramaturgy in the American civil rights movement' in McAdam et al. (1996), 338–355.

- McAdam, Doug; John D. McCarthy and Mayer N. Zald, eds. (1996) *Comparative perspectives on social movements. Political opportunities, mobilizing structures, and cultural framings*. Cambridge University Press.
- McCall, Catherine (1990) *Concepts of Person. An analysis of concepts of person, self and human being*. Avebury.
- McConahay, John B.; Hardee, Betty B. and Batts, Valerie (1981) 'Has Racism Declined in America? It Depends on Who Is Asking and What Is Asked' *The Journal of Conflict Resolution*, 25:4, 563–579.
- McDowell, John (1996) *Mind and World*. Cambridge.
- (2002) 'Responses' in Smith (2002), Chapter 14, 269–305.
- McGuire, Danielle L. (2010) *At the Dark End of the Street. Black Women, Rape, and Resistance. A New History of the Civil Rights Movement from Rosa Parks to the Rise of Black Power*. Vintage Books.
- McHugh, Conor (2013) 'Normativism and Doxastic Deliberation' *Analytic Philosophy*, 54:4, 447–465.
- McKenna, Michael (2005) 'The Relationship between Autonomous and Morally Responsible Agency' in James Stacey Taylor, ed. (2005) *Personal Autonomy. New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, 1–29. Cambridge University Press.
- (2012) 'Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism' *Ethics*, 16, 145–174.
- (2013) 'Reasons-Responsiveness, Agents, and Mechanisms' in Shoemaker (2013), Chapter 6, 151–183.
- McKenna, Michael and David Widerker (2003) 'Introduction' in Widerker and McKenna (2003), 1–16.
- McKenna, Michael and Coates, D. Justin (2015) 'Compatibilism' *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zalta, ed. URL = <<http://plato.stanford.edu/archives/sum2015/entries/compatibilism/>>.
- McKeon, Richard (1957) 'The Development and the Significance of the Concept of Responsibility' in McKeon (1990), Chapter 3, 62–87.
- (1990) *Freedom and History and Other Essays*.
- McReynolds, Paul (1980) 'The clock metaphor in the history of psychology' in Nickles (1980), 97–112.
- McQuillan, Martin (2000) 'Introduction. Aporias of Writing: Narrative and Subjectivity' in Martin McQuillan, ed. (2000) *The Narrative Reader*, 1–33. Routledge.
- Medina, José (2013) *The Epistemology of Resistance. Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford University Press.
- Mele, Alfred R. (1995) *Autonomous Agents. From Self-Control to Autonomy*. Oxford University Press.
- Melley, Timothy (2008) 'Brainwashed! Conspiracy Theory and Ideology in the Postwar

- United States' *New German Critique*, 103, Dark Powers: Conspiracies and Conspiracy Theory in History and Literature, 145–164.
- Mendus, Susan (1986 – 1987) 'Liberty and Autonomy' *Proceedings of the Aristotelian Society*, New Series, 87, 107–120.
- (1992) 'Strangers and Brothers: Liberalism, Socialism and the Concept of Autonomy' in Milligan and Miller (1992), 3–16.
- Merzat-Bruns, Marie (2013/ 2016) *Discrimination at Work. Comparing European, French, and American Law*, translated by Elaine Holt (original title: *Discriminations en droit du travail. Dialogue avec la doctrine américaine*). Oakland, California: University of California Press.
- Metzinger, Thomas (2009) *The Ego Tunnel. The Science of the Mind and the Myth of the Self*. Basic Books.
- Meyers, Diana Tietjens (2005) 'Decentralizing Autonomy: Five Faces of Selfhood' in Christman and Anderson (2005), Chapter 2, 27–55.
- Mickelson, Kristin (2017) 'The Manipulation Argument' in Timpe, Griffith and Levy (2017), Chapter 14, 166–178.
- Milanese, Arnaud (2014) 'Nécessité et imputation chez Hobbes. Se démarquer d'Aristote et se démarquer de la scolastique' *Philosophiques*, 41:1, 3–35.
- Mill, John Stuart (1865 and 1872/ 1979) *An Examination of Sir William Hamilton's Philosophy. Collected Works of John Stuart Mill. Volume IX*. Routledge.
- Mills, Charles W. (1997) *The Racial Contract*. Ithaca and London: Cornell University Press.
- Miller, Cecilia (1993) *Giambattista Vico: Imagination and Historical Knowledge*. The MacMillan Press Ltd.
- Milligan, David and William Watts Miller, eds. (1992) *Liberalism, Citizenship and Autonomy*. Avebury.
- Millgram, Elijah, ed. (2001) *Varieties of Practical Reasoning*. The MIT Press.
- Mills, Charles W. (1997) *The Racial Contract*. Cornell University Press.
- Mills, Jon and Polanowski, Janusz (1997) *The Ontology of Prejudice*. Rodopi.
- Minow, Martha (2010) *In Brown's Wake. Legacies of America's Educational Landmark*. Oxford University Press.
- Misak, Cheryl (2000) *Truth, Politics, Morality. Pragmatism and deliberation*. Routledge.
- Mittelstraß, Jürgen, ed. (1984/ 1995) *Enzyklopädie Philosophie und Wissenschaftstheorie*. J.B. Metzler.
- (1995) 'Mentalismus', in Mittelstraß (1984/ 1995), Part 2, H-O, 847–848.
- Mohyeldin Said, K.A.; Newton-Smith, W.H.; Viale, R. and Wilkes, K.V., eds (1990), *Modelling the Mind*. Clarendon Press.
- Mölder, Bruno (2010) *Mind Ascribed. An elaboration and defence of interpretivism*. John Benjamins Publishing Company.
- Moody-Adams, Michele M. (1994) 'Culture, Responsibility, and Affected Ignorance'

- Ethics, 104:2, 291–309.
- Moore, Barrington, jr. (1978) *The Social Bases of Obedience and Revolt*. M.E. Sharpe.
- Moore, J. (2005) ‘Some Historical and Conceptual Background to the Development of B.F. Skinner’s “Radical Behaviorism” – Part 2’ *The Journal of Mind and Behavior*, 26:1/2, *Special Section: When Did B.F. Skinner Become a Radical Behaviorist?*, 95–123.
- Moors, Agnes and Jan De Houwer (2006) ‘Automaticity: A Theoretical and Conceptual Analysis’ *Psychological Bulletin*, 132:2, 297–326.
- Moran, Dermot, ed. (2008) *The Routledge Companion to Twentieth Century Philosophy*. Routledge.
- Moran, Richard (2001) *Authority and Estrangement. An Essay on Self-Knowledge*. Princeton University Press.
- (2004) ‘Replies to Heal, Reginster, Wilson, and Lear’. *Philosophy and Phenomenological Research*, 69:2, 455–472.
- Morris, Christopher W. and Ripstein, Arthur, eds. (2001) *Practical Rationality and Preference. Essays for David Gauthier*. Cambridge University Press.
- Mortati, Costantino and Santoro-Passarelli, Francesco, eds. (1983) *Enciclopedia del diritto. Volume XXXIII. Perenzionem – Pluralismo*. Giuffrè Editore.
- Morton, Adam (1990) ‘Why there is no Concept of a Person’ in Gill (1990), Chapter 1, 39–59.
- Moxley, Roy A. (2004) ‘B.F. Skinner’s Adoption of Peirce’s Pragmatic Meaning for Habits’ *Transactions of the Charles S. Peirce Society*, 40:4, 743–769.
- Moyal-Sharrock, Danièle (2005) *Understanding Wittgenstein’s On Certainty*. Palgrave Macmillan.
- Mouffe, Chantal (2005) ‘The Limits of John Rawls’s pluralism’ *Politics, Philosophy & Economics*, 4, 221–231.
- Mulhall, Stephen and Swift, Adam (1992/ 1996) *Liberals & Communitarians*. Second Edition. Blackwell Publishing.
- Muller, John and Brent, Joseph, eds. (2000) *Peirce, Semiotics, and Psychoanalysis*. The John Hopkins University Press.
- Murray, J.A.H., ed. (1909) *A New English Dictionary on Historical Principles; founded mainly on the materials collected by The Philological Society. Volume VII. O, P. Part II, PH to PY*. At the Clarendon Press.
- Murray, John B. (2016) ‘Hypnosis and Criminal Behavior’ *The Catholic Lawyer*, 11:3, 209–214.
- Myrdal, Gunnar (1944/ 1962) *An American Dilemma. The Negro Problem & Modern Democracy. Volume 1*. Pantheon Books.
- (1944/ 1962) *An American Dilemma. The Negro Problem & Modern Democracy. Volume 2*. Pantheon Books.
- Nanay, Bence (forthcoming) ‘Implicit bias as mental imagery’ *Journal of the American*

Philosophical Association.

- Neely, Wright (1974) 'Freedom and Desire' *Philosophical Review*, 83, 32–54.
- Nehamas, Alexander (1981) 'The Postulated Author: Critical Monism as a Regulative Ideal' *Critical Inquiry*, 8:1, 133–149.
- Nelkin, Dana Kay (2011) *Making Sense of Freedom and Responsibility*. Oxford University Press.
- Neumann, Michael (2000) 'Did Kant Respect Persons?' *Res Publica*, 6, 285–299.
- Ngo, Helen (2016) 'Racist habits: A phenomenological analysis of racism and the habitual body' *Philosophy and Social Criticism*, 42(9), 847–872.
- (2017) *The Habits of Racism. A Phenomenology of Racism and Racialized Embodiment*. Lexington Books.
- Nichols, Shaun and Stephen P. Stich (2003) *Mindreading: An Integrated Account of Pretence, Self-awareness, and Understanding Other Minds*. Oxford University Press.
- Nicholls, Angus and Martin Liebscher (2010a) 'Introduction: thinking the unconscious' in Nicholls and Liebscher (2010b), 1–25.
- eds. (2010b) *Thinking the Unconscious. Nineteenth-Century German Thought*. Cambridge University Press.
- Nickles, Thomas, ed. (1980) *Scientific Discovery: Case Studies*. D. Reidel Publishing Company.
- Niebuhr, H. Richard (1963/ 1999) *The Responsible Self: An Essay in Christian Moral Philosophy*. Westminster John Knox Press.
- Nier, Jason A., ed. (2013) *Taking Sides. Clashing Views in Social Psychology*. Fourth Edition. The McGraw Hill Companies.
- Nisbett, Richard E. and Wilson, Timothy DeCamp (1977) 'Telling More Than We Can Know: Verbal Reports on Mental Processes' *Psychological Review*, 84:3, 231–259.
- Noggle, Robert (1996) 'Manipulative Actions: A Conceptual and Moral Analysis' *American Philosophical Quarterly*, 33:1, 43 – 55.
- (2018) 'The Ethics of Manipulation' *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/sum2018/entries/ethics-manipulation/>>.
- Norval, Aletta J. (2007) *Aversive Democracy. Inheritance and Originality in the Democratic Tradition*. Cambridge University Press.
- Nosek, Brian A. and Hansen, Jeffrey J. (2008) 'The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation' *Cognition and Emotion*, 22:4, 553–594.
- Normore, Calvin G. (2002) 'Goodness and Rational Choice in the Early Middle Ages' in H. Lagerlund and M. Yrjönsuuri, eds., *Emotions and Choice from Boethius to Descartes*, 29–47. Kluwer Academic Publishers.
- (2007) 'Freedom, Contingency, and Rational Power' *Proceedings and Addresses of the American Philosophical Association*, 81:2, 49–64.

- Nottelmann, Nikolay, ed. (2013) *New Essays on Belief: Constitution, Content and Structure*. Palgrave Macmillan
- Nuttall, A.D. (1978) *Dostoevsky's Crime and Punishment. Murder as Philosophic Experiment*. Sussex University Press.
- Nuzzo, Angelica (1994) 'Metamorphosen der Freiheit in der Jenenser Kant-Rezeption (1785–1794)' in Strack (1994), 484–518.
- (2010) 'Analisi Filosofica e Coscienza Storica: Kant e Hegel oggi' *Studi Kantiani*, 23, 77–87.
- Nyquist (2009) 'Hobbes, Slavery, and Despotical Rule'.
- Obasogie, Osagie K. (2010) 'Do Blind People See Race? Social, Legal, and Theoretical Considerations' *Law & Society Review*, 44:3/4, 585–616.
- O'Connor, Timothy (2000) *Persons and Causes. The Metaphysics of Free Will*. Oxford University Press.
- Offord, Derek (1983) 'The causes of crime and the meaning of law: Crime and Punishment and contemporary radical thought' in Jones and Terry (1983), Chapter 2, 41–65.
- O'Neill, Onora (2000/ 2012) 'Kant and the Social Contract Tradition' in Ellis (2012), Chapter 1, 25–41.
- (2003) 'The Inaugural Address: Autonomy: The Emperor's New Clothes' *Proceedings of the Aristotelian Society, Supplementary Volumes*, 77, 1–21.
- Olson, Micheal A. and Russell H. Fazio (2002) 'Implicit acquisition and manifestation of classically conditioned attitude' *Social Cognition*, 20:2, 89–103.
- Open Society Foundations (2016) *Strategic Litigation Impacts. Roma School Desegregation*. <https://eric.ed.gov/?id=ED606038>.
- Orfield, Gary (2001) *Schools More Separate: Consequences of a Decade of Segregation*. Harvard University Project.
- Orne, Martin T. (1972) 'Can a hypnotized subject be compelled to carry out otherwise unacceptable behavior? A discussion' *International Journal of Clinical and Experimental Hypnosis*, 20:2, 101–117.
- Orosz, Gábor; Erzsébet Bánki; Béata Böthe; István Tóth-Király; Linda R. Tropp (2016) 'Don't judge a living book by its cover: effectiveness of the living library intervention in reducing prejudice toward Roma and LGBT people' *Journal of Applied Social Psychology*, 46, 510–517.
- Oshana, Marina A. (2002) 'The Misguided Marriage of Responsibility and Autonomy' *The Journal of Ethics*, 6, 261–280.
- Oskamp, Stuart, ed. (2000) *Reducing Prejudice and Discrimination*. Psychology Press.
- Ostrow, James M. (1981) 'Culture as a Fundamental Dimension of Experience: A Discussion of Pierre Bourdieu's Theory of Human Habitus' *Human Studies*, 4:3, 279–297.
- Ottmann, Henning (1981) 'Herr und Knecht bei Hegel. Bemerkungen zu einer

- mißverstandenen Dialektik' *Zeitschrift für philosophische Forschung*, 35:3/4, 365–384.
- Overgaard, Morten (2006) 'Introspection in Science' *Consciousness and Cognition*, 15, 629–633.
- Overton, Willis F. and Palermo, David S. (1994) *The Nature and Ontogenesis of Meaning*. Lawrence Erlbaum Associates, Publishers.
- Owens, David (2000) *Reason without Freedom. The problem of epistemic normativity*. Routledge.
- (2003) 'Does belief have an aim?' *Philosophical Studies*, 115:3, 283–305.
- Oxford English Dictionary (2000) 'Manipulation' *OED Third Edition*, September. Oxford University Press.
- Packard, Vance (1957/ 2007) *The Hidden Persuaders*. Ig Publishing.
- Pager, Devah and Shepherd, Hana (2008) 'The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets' *Annual Review of Sociology*, 1:34, 181–209.
- Parfit, Derek (1984) *Reasons and Persons*. Clarendon Press.
- Parker, Andrew; Sarat, Austin and Umphrey, Martha Merrill, eds. (2011) *Subjects of Responsibility. Framing Personhood in Modern Bureaucracies*. Fordham University Press.
- Pasnau, Robert and Van Dyke, Christina, eds. (2014) *The Cambridge History of Medieval Philosophy. Volume I. Revised Edition*. Cambridge University Press.
- Pateman, Carole (1988) *The Sexual Contract*. Polity Press.
- Paul, Ellen Frankel, Fred D. Miller, Jr., and Jeffrey Paul, eds. (2003) *Autonomy*. Cambridge University Press.
- Payne, B. Keith and Bertram Gawronski (2010) 'A History of Implicit Social Cognition. Where Is It Coming From? Where Is It Now? Where Is It Going?' in Gawronski and Payne (2010), Chapter 1, 1–15.
- Payrató, Lluís (2009) 'Non-verbal communication' in Jef Verschueren and Jan-Ola Östman, eds. (2009) *Key Notions for Pragmatics*, 163–194. John Benjamins Publishing Company
- Peacocke, Christopher (2005) 'Joint Attention: Its Nature, Reflexivity, and Relation to Common Knowledge' in Eilan; Hoerl; McCormack, and Roessler (2005), Chapter 14, 298–324.
- Peels, Rik and Blaauw, Martijn (2016) *The Epistemic Dimensions of Ignorance*. Cambridge University Press.
- Pehrson, Samuel and Leach, Colin Wayne (2012) 'Beyond 'old' and 'new': for a social psychology of racism' in Dixon and Levine (2012), Chapter 6, 120–138.
- Peirce, Charles (1866) 'Lowell Lecture VI' in Peirce (1982), 440–454.
- (1982) *Writings of Charles S. Peirce. A chronological edition. Volume I. 1857–1866*. Indiana University Press.

- Pereboom, Derk (1995) 'Determinism al Dente' *Noûs*, 29:1, 21–45.
- (2001) *Living Without Free Will*. Cambridge University Press.
- (2004) 'Is Our Conception of Agent-Causation Coherent?'
- ed. (2009a) *Free Will*. Second Edition. Hackett Publishing Company.
- (2009b) 'Introduction' in Pereboom (2009a).
- (2013) 'Optimistic Skepticism about Free Will' in Russell and Deery (2013), Chapter 22, 421–449.
- Peregrin, Jaroslav (2014) *Inferentialism. Why Rules Matter*. Palgrave Macmillan.
- Pérez-Ramos, Antonio (1996) 'Giambattista Vico' in Brown (1996), Chapter 13, 273–292.
- Pervin, Lawrence A. and John, Oliver P. (1999) *Handbook of personality. Theory and Research*. Second Edition. The Guilford Press.
- Pettigrew, Thomas F. (1979) 'Racial Change and Social Policy' *The ANNALS of the American Academy of Political and Social Science*, 441:1, 114–131.
- (2008) 'Still a Long Way To Go: American Black–White Relations Today' in Adams et al. (2008), 45–61.
- Pettigrew, Thomas F. and Meertens, R.W. (1995) 'Subtle and blatant prejudice in western Europe' *European Journal of Social Psychology*, 25, 57–75.
- Pettit, Philip (2008) *Made with Words. Hobbes on Language, Mind, and Politics*. Princeton University Press.
- Pettit, Philip and Michael Smith (1996) 'Freedom in Belief and Desire' *The Journal of Philosophy*, 93, 429–449.
- Phillips, Anne (2013) *Our Bodies, Whose Property?* Princeton University Press.
- Pickel, Andreas (2005) 'The Habitus Process: A Biopsychosocial Conception' *Journal for the Theory of Social Behaviour*, 35:4, 437–461.
- Pincus, Fred L. (1999) 'From Individual to Structural Discrimination' in Pincus and Ehrlich (1999), 120–124.
- Pincus, Fred L. and Ehrlich, Howard J. (1999) *Race and Ethnic Conflict. Contending Views on Prejudice, Discrimination, and Ethnviolence*. Second edition. Westview.
- Pippin, Robert (2000) 'Hegel's practical philosophy: the realization of freedom' in Ameriks (2000), Chapter 9, 180–199.
- (2007) 'Brandom's Hegel' in Hammer (2007), Chapter 7, 153–180.
- Plant, E. Ashby and B. Michelle Peruche (2005) 'The Consequences of Race for Police Officers' Responses to Criminal Suspects' *American Psychological Society*, 16:3, 180–183.
- Plessy v. Ferguson*, 163 U.S. 537 (1896).
- Pockett, Susan; William P. Banks and Shaun Gallagher, eds. (2006) *Does Consciousness Cause Behavior?* The MIT Press.
- Pojman, Louis P. and Owen McLeod, eds. (1999) *What Do We Deserve? A Reader on Justice and Desert*. Oxford University Press.

- Pollard, Bill (2008) *Habits in Action. A Corrective to the Neglect of Habits in Contemporary Philosophy of Action*. VDM Verlag Dr. Müller.
- Polletta, Francesca (2002) 'Plotting Protest: Mobilizing Stories in the 1960 Student Sit-Ins' in Davis (2002a), 31–51.
- (2006) *It Was Like a Fever. Storytelling in Protest and Politics*. The University of Chicago Press.
- Polletta, Francesca; Pang Ching Bobby Chen; Beth Gharrity Gardner and Alice Motes (2011) 'The Sociology of Storytelling' *Annual Review of Sociology*, 37, 109–130.
- Pöltner, Günther, ed. (1981) *Personale Freiheit und pluralistische Gesellschaft*. Herder.
- Pool, Heather (2012) 'Mourning Emmett Till' *Law, Culture and the Humanities*, 1–31.
- Popat, Shreeya and William Winslade (2015) 'While You Were Sleepwalking: Science and Neurobiology of Sleep Disorders & the Enigma of Legal Responsibility of Violence During Parasomnia' *Neuroethics*, 8, 203–214.
- Posner, Richard A. (2003) *Law, Pragmatism, and Democracy*. Harvard University Press.
- Post, Robert C. (1998a) 'Censorship and Silencing' in Post (1998b), Chapter 1, 1–12.
- ed. (1998b) *Censorship and Silencing: Practices of Cultural Regulation*. The Getty Research Institute.
- Pound, Roscoe (1954) 'The Role of the Will in Law' *Harvard Law Review*, 68:1, 1–19.
- Pratkanis, Anthony R. (1992) 'The Cargo-Cult Science of Subliminal Persuasion' in Nier (2013), 301–312.
- Price, Huw (2008) 'Will There Be Blood? Brandom, Hume, and the Genealogy of Modals' *Philosophical Topics*, 36:2, 87–97.
- Prien, Bernd and Schweikard, David P., eds. (2008) *Robert Brandom. Analytic Pragmatist*. Ontos Verlag.
- Pritzlaff, Tanja (2008) 'Freedom is a Matter of Responsibility and Authority: An Interview with Robert B. Brandom' *European Journal of Political Theory*, 7, 365–381.
- Public Defender of Rights of the Czech Republic (2019) *Opinion of the Public Defender of Rights on the Communication from the Czech authorities concerning enforcement of the judgement of the European Court of Human Rights in the Case of D. H. and Others v. the Czech Republic*, accompanied with a letter to the Department for the Execution of Judgments of the ECHR, published 14 May 2019. Available at https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=090000168094d830 (accessed on 25 March 2021).
- Pufendorf, Samuel (1660/ 1999) *Elementa jurisprudentiae universalis*. Akademie Verlag.
- Quillian, Lincoln (2006) 'New Approaches to Understanding Racial Prejudice and Discrimination' *Annual Review of Sociology*, 32, 299–328.
- Quillian, Lincoln; Devah Pager; Ole Hexel and Arnfinn H. Midtbøen (2017) 'Meta-analysis of field experiments shows no change in racial discrimination in hiring over time' *Proceedings of the National Academy of Sciences of the United States of*

- America*, 114:41, 10870 – 10875.
- Rabb, J. Douglas (1977) 'Reflection, Reflexion and Introspection' *The Locke newsletter*, 8, 35-52.
- Rabinow, Paul (1986) 'Representations Are Social Facts: Modernity and Post-Modernity in Anthropology' in Paul Rabinow (1996) *Essays on the Anthropology of Reason*, Chapter 2, 28–58. Princeton University Press.
- Railton, Peter (1997) 'On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action' in Cullity and Gaut (1997), Chapter 2, 53–79.
- Ranouil, Véronique (1980) *L'Autonomie de la Volonté. Naissance et évolution d'un concept*. Presses Universitaires de France.
- Ranson, Ina (2008) "'Verantwortung" in Germany. Encountering the Sense of Responsibility' in Sizoo (2008), Chapter 10, 261–285.
- Rawls, John (1971) *A Theory of Justice*. The Belknap Press of Harvard University Press.
- (1993/ 1996) *Political Liberalism*. Columbia University Press.
- (1969a) 'The Justification of Civil Disobedience' in Rawls (1999b), 176–189.
- (1969b) 'A Kantian Conception of Equality' in Rawls (1999b), 255–266.
- (1980) 'Kantian Constructivism in Moral Theory' *The Journal of Philosophy*, 77:9, 515–572.
- (1985) 'Justice as Fairness: Political not Metaphysical' *Philosophy & Public Affairs*, 14:3, 223–251.
- (1999a) *A Theory of Justice. Revised Edition*. The Belknap Press of Harvard University Press.
- (1999b) *Collected Papers*. Harvard University Press
- (2001/ 2011) *Justice as Fairness. A Restatement*. Universal Law Publishing.
- Raz, Joseph (1986) *The Morality of Freedom*. Clarendon Press.
- Reid, Thomas (1793) 'On a novel use of the word motive – Causality of motives, &c.' in Hamilton (1872), 87–88.
- Reidy, David A. and Riker, Walter J., eds. (2010) *Coercion and the State*. Springer.
- Reisinger, K. and O.R. Scholz (2019) 'Vorurteil' in Ritter et al. (1971 – 2007), Volume 11, 1250–1263.
- Resnik, Judith and Curtis, Dennis (2011) *Representing Justice. Invention, Controversy, and Rights in City-States and Democratic Courtrooms*. Yale University Press.
- Ricciardi, Mario (2007) 'Berlin on Liberty' in Crowder (2005), Chapter 5, 119–139.
- Richards, Graham (1992) *Mental Machinery. The Origins and Consequences of Psychological Ideas. Part 1: 1600–1850*. The John Hopkins University Press.
- Rieber, Robert W. (2006) *The Bifurcation of the Self. The History and Theory of Dissociation and Its Disorders*. Springer.
- Ritter, Joachim and Karlfried Gründer, eds. (1971–2007/ 2019a) *Historisches Wörterbuch der Philosophie. Volume 5: L–Mn*. Wissenschaftliche Buchgesellschaft.
- eds. (1971–2007/ 2019b) *Historisches Wörterbuch der Philosophie. Volume 9: Se –*

- Sp.* Wissenschaftliche Buchgesellschaft.
- Ritter, Joachim; Karlfried Gründer and Gottfried Gabriel, eds. (1971–2007/ 2019) *Historisches Wörterbuch der Philosophie. Volume 11: U–V*. Wissenschaftliche Buchgesellschaft.
- Roberts, Gene and Hank Klibanoff (2006) *The Race Beat: The Press, the Civil Rights Struggle, and the Awakening of a Nation*. Vintage Books.
- Roberts, Rodney C. (2002a) ‘Justice and Rectification: A Taxonomy of Justice’ in Roberts (2002), 7–28.
- ed. (2002b) *Injustice and Rectification*. Peter Lang.
- Robeyns, Ingrid (2008) ‘Ideal Theory in Theory and Practice’ *Social Theory and Practice*, 34:3, 341–362.
- Roediger, Henry L., III; Michael K. Goode and Franklin M. Zaromb (2008) ‘Free Will and the Control of Action’ in Baer, Kaufman and Baumeister (2008), Chapter 10, 205–225.
- Rooth, Dan-Olof (2007) ‘Implicit discrimination in hiring: real world evidence’, IZA Discussion Papers, No. 2764, Institute for the Study of Labor (IZA), Bonn, <http://nbn-resolving.de/urn:nbn:de:101:1-20080402107>.
- Rorty, Amélie Oksenberg, ed. (1980) *Essays on Aristotle’s Ethics*. University of California Press.
- Rorty, Richard (1979/ 2009) *Philosophy and the Mirror of Nature. Thirtieth-Anniversary Edition*. Princeton University Press.
- Rosanvallon, Pierre (1995) *La nouvelle question sociale*. Éditions du Seuil.
- Rose, Reginald (1955/ 2017) *Twelve Angry Men*. Bloomsbury.
- Rosen, Michael (2012) Dignity. Its History and Meaning.
- Rosenthal, David M. (1999) ‘Introspection’ in Robert A. Wilson and Frank C. Keil (1999), 419–420.
- Roth, Andreas (2008) ‘Crimen contra naturam’ in Daston and Stolleis (2008b), 89–103.
- Rothstein, Richard (2017) *The Color of Law: A Forgotten History of How Our Government Segregated America*. New York and London: Liveright Publishing Corporation.
- Rowe, Mary (1990) ‘Barriers to Equality: The Power of Subtle Discrimination to Maintain Unequal Opportunity’ *Employee Responsibilities and Rights Journal*, 3:2, 153–163.
- (2008) ‘Micro-affirmations & Micro-inequities’ *Journal of the International Ombudsman Association*, 1:1, 45–48.
- Rudinow, Joel (1978) ‘Manipulation’ *Ethics*, 88:4, 338–347.
- Rudman, Laurie A. (2004a) ‘Sources of Implicit Attitudes’ *Current Directions in Psychological Science*, 13:2, 79–82.
- (2004b) ‘Social Justice in Our Minds, Homes, and Society: The Nature, Causes, and Consequences of Implicit Bias’ *Social Justice Research*, 17:2, 129–142.
- Rudman, Laurie A. and Glick, Peter (2001) ‘Prescriptive Gender Stereotypes and

- Backlash Toward Agentic Women' *Journal of Social Issues*, 57:4, 743–762.
- Ruggiero, Guido De (1925/ 2003) *Storia del liberalismo europeo*. Editori Laterza.
- Rush, Fred and Stolzenberg, Jürgen (2013) *Internationales Jahrbuch des Deutschen Idealismus/ International Yearbook of German Idealism. Freiheit/ Freedom*. Walter de Gruyter.
- Russell, Paul (1995) *Freedom and Moral Responsibility. Hume's Way of Naturalizing Responsibility*. Oxford University Press.
- Russell, Paul and Oisín Deery, eds. (2013) *The Philosophy of Free Will*. Oxford University Press.
- Růžička, Adam (2014) *Implicit prejudice: The Role of Incidental Disgust and Anger*. Bachelor thesis (unpublished).
- Ryle, Gilbert (1949/ 2009) *The Concept of Mind*. Routledge.
- Rysiew, Patrick (2008) 'Rationality Disputes – Psychology and Epistemology' *Philosophy Compass*, 3:6, 1153–1176.
- Sabl, Andrew (2001) 'Looking Forward to Justice: Rawlsian Civil Disobedience and its Non-Rawlsian Lessons' *The Journal of Political Philosophy*, 9:3, 307–330.
- Saine, Thomas P. (1997) *The Problem of Being Modern. Or The German Pursuit of Enlightenment from Leibniz to the French Revolution*.
- Salter, Phia and Adams, Glenn (2013) 'Toward a Critical Race Psychology' *Social and Personality Psychology Compass*, 7:11, 781–793.
- Sandel, Michael J. (1982/ 1998) *Liberalism and the Limits of Justice*. Second Edition. Cambridge University Press.
- Sanders, Lynn M. (1997) 'Against Deliberation' *Political Theory*, 25:3, 347–376.
- Sandywell, Barry (1996) *Reflexivity and the Crisis of Western Reason. Logological Investigations. Volume 1*. Routledge.
- (2004) 'The myth of everyday life: Toward a heterology of the ordinary' *Cultural Studies*, 18:2/3, 160–180.
- Santoro, Emilio (2003) *Autonomy, Freedom and Rights. A Critique of Liberal Subjectivity*. Kluwer Academic Publishers. Original publication in Italian: (1999) *Autonomia individuale, libertà e diritti. Una critica dell' antropologia liberale*. Edizioni Ets.
- Sarasohn, Lisa T. (1985) 'Motion and Morality: Pierre Gassendi, Thomas Hobbes and the Mechanical World-View' *Journal of the History of Ideas*, 46:3, 363–379.
- Sarat, Austin, ed. (2015) *The Punitive Imagination. Law, Justice, and Responsibility*. The University of Alabama Press.
- Sartre, Jean-Paul (1943) *L'être et le néant. Essai d'ontologie phénoménologique*. Éditions Gallimard.
- Sass, Louis A. (1994) *The Paradoxes of Delusion. Wittgenstein, Schreber, and the Schizophrenic Mind*. Cornell University Press.
- Saul, Jennifer (2013) 'Implicit Bias, stereotype threat, and women in philosophy' in Hutchison and Jenkins (2013), Chapter 2, 39–60.

- (2017) ‘Implicit Bias, Stereotype Threat, and Epistemic Injustice’ in Kidd et al. (2017), Chapter 8, 235–242.
- Scanlon, T.M. (1986) ‘The Significance of Choice’ *The Tanner Lectures on Human Values*, 7: 149–216.
- (1998) *What We Owe to Each Other*. The Belknap Press of Harvard University Press.
- Scheffler, Samuel (1992) ‘Responsibility, Reactive Attitudes, and Liberalism in Philosophy and Politics’ *Philosophy & Public Affairs*, 21:4, 299–323.
- Schlick, Moritz (1930/ 1966) ‘When Is A Man Responsible?’ (original title: ‘Wann ist der Mensch verantwortlich?’), in Berofsky (1966), 54–63.
- Schmid, Hans Bernhard and Schweikard, David P. (2009a) ‘Einleitung: Kollektive Intentionalität. Begriff, Geschichte, Probleme’ in Schmid and Schweikard (2009b), 11–65.
- eds. (2009b) *Kollektive Intentionalität. Eine Debatte über die Grundlagen des Sozialen*. Suhrkamp Verlag.
- Schmidt-Degenhard, Tobias Joachim (2008) *Robert Ritter (1901–1951). Zu Leben und Werk des NS-“Zigeunerforschers”*. Dissertation (unpublished).
- Schnädelbach, Herbert (1977) *Reflexion und Diskurs. Fragen einer Logik der Philosophie*. Suhrkamp Verlag.
- (1992a) *Zur Rehabilitierung des animal rationale. Vorträge und Abhandlungen 2*. Suhrkamp.
- (1992b) ‘Rationalität und Normativität’ in Schnädelbach (1992a), 79–103.
- (2000a) *Philosophie in der modernen Kultur*. Suhrkamp.
- (2000b) ‘Rationalitätstypen’ in Schnädelbach (1992a), 256–281.
- Schneewind, Jerome B. (1987) ‘Pufendorf’s Place in the History of Ethics’ *Synthese*, 72(1), 123–155.
- (1996) ‘Voluntarism and the Foundations of Ethics’ *Proceedings and Addresses of the American Philosophical Association*, 70:2, 25–41.
- (1998) *The Invention of Autonomy. A History of Modern Moral Philosophy*. Cambridge University Press.
- (2013) ‘Autonomy after Kant’ in Sensen (2013), Chapter 8, 146–168.
- Schneider, Hans Julius (1975) *Pragmatik als Basis von Semantik und Syntax*. Suhrkamp Verlag.
- Schneiders, Werner (1983) *Aufklärung und Vorurteilkritik. Studien zur Geschichte der Vorurteilttheorie*. Frommann-holzboog.
- Schoeman, Ferdinand, ed. (1987) *Responsibility, Character, and the Emotions. New Essays in Moral Psychology*. Cambridge University Press.
- Schönecker, Dieter (2013) ‘“A free will and a will under moral laws are the same”: Kant’s concept of autonomy and his thesis of analyticity in Groundwork III’ in Sensen (2013), Chapter 12, 225–245.

- Schrager, Cynthia D. (1996) 'Both Sides of the Veil: Race, Science, and Mysticism in W.E.B. Du Bois' *American Quarterly*, 48:4, 551–586.
- Schröder, Peter (2018) 'Fidem observandam esse – Trust and Fear in Hobbes and Locke' in Kontler and Somos (2018), Chapter 6, 99–117.
- Schwartz, Marlene B.; Chambliss, Heather O'Neal; Brownell, Kelly D.; Blair, Steven N. and Billington, Charles (2003) 'Weight Bias among Health Professionals Specializing in Obesity' *Obesity Research*, 11(9), 1033–1039.
- Schweikard, David P. and Hans Bernhard Schmid (2013) 'Collective Intentionality' *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/sum2013/entries/collective-intentionality/>>.
- Schwitzgebel, Eric (2010) Acting Contrary to Our Professed Beliefs or the Gulf Between Occurrent Judgment and Dispositional Belief' *Pacific Philosophical Quarterly*, 91, 531–553.
- Scott, James (1990) *Domination and the Arts of Resistance*. Hidden Transcripts. Yale University Press.
- Searle, John R. (1979) 'What Is an Intentional State?' *Mind*, New Series, 88:349, 74 – 92.
- (1983) *Intentionality. An essay in the philosophy of mind*. Cambridge University Press.
- (1990) 'Collective Intentions and Actions' in Searle (2002), 90 – 105.
- (1995) *The Construction of Social Reality*. The Penguin Press.
- (2002) *Consciousness and Language*. Cambridge University Press.
- Seel, Martin (2005) 'Teilnahme und Beobachtung. Zu den Grundlagen der Freiheit' *Neue Rundschau*, 116:4, 141–153.
- Seidler, Youri; Romy van Leeuwen; L.A.J. Koster; Marion van San; Paul van Wensveen; Peter Jorna (2020) *Monitor Sociale Inclusie (meting 4). Derde vervolgmeting naar de woon- en leefomstandigheden van Roma en Sinti in Nederland*. Rotterdam: Erasmus Universiteit Rotterdam/ Risbo.
- Selisker, Scott (2016) *Human Programming. Brainwashing, Automaton, and American Unfreedom*. University of Minnesota Press.
- Sellaro, Roberta; Derks, Belle; Nitsche, Michael A.; Hommel, Bernhard; van den Wildenberg, Wery P.M.; van Dam, Kristina and Colzato, Lorenza S. (2015) 'Reducing Prejudice Through Brain Stimulation' *Brain Stimulation*, 8, 891–897.
- Selmi, Michael (2018) 'The Paradox of Implicit Bias and a Plea for a New Narrative' *Arizona State Law Journal*, 50:1, 193–246.
- Sensen, Oliver, ed. (2013) *Kant on Moral Autonomy*. Cambridge University Press.
- Sgarbi, Marco (2012) *Kant on Spontaneity*. Bloomsbury.
- Shah, Nishi (2002) 'Clearing Space for Doxastic Voluntarism' *The Monist*, 85:3, Controlling Belief, 436–445.

- (2003) ‘How Truth Governs Belief’ *The Philosophical Review*, 112:4, 447–482.
- Shah, Timothy Samuel (2000) ‘Making the Christian World Safe for Liberalism: From Grotius to Rawls’ *The Political Quarterly*, 71, 121-139.
- Shalamov, Varlam (1980/ 1994) *Kolyma Tales*. Penguin Books.
- Sharp, Joanne P.; Routledge, Paul; Philo, Chris and Paddison, Ronan (2000a) ‘Entanglements of Power. Geographies of domination/resistance’ in Sharp, Routledge, Philo and Paddison (200b), Chapter 1, 1–42.
- eds. (2000b) *Entanglements of power. Geographies of domination/resistance*. Routledge.
- Sher, George (1987) *Desert*. Princeton University Press.
- (2009) *Who Knew? Responsibility Without Awareness*. Oxford University Press.
- Sherman, Jeffrey W.; Bertram Gawronski and Yaacov Trope, eds. (2014) *Dual-Process Theories of the Social Mind*. The Guilford Press.
- Shoemaker, David, ed. (2013) *Oxford Studies in Agency and Responsibility, Volume 1*. Oxford University Press.
- Shotwell, Alexis (2011) *Knowing Otherwise. Race, Gender, and Implicit Understanding*. The Pennsylvania State University Press.
- Sie, Maureen and Nicole van Voorst Vader-Bours (2016) ‘Stereotypes and Prejudices: Whose Responsibility? Indirect Personal Responsibility for Implicit Biases’ in Brownstein and Saul (2016), Chapter 1.4, 90–114.
- Sieckmann, Jan-R (2012) *The Logic of Autonomy. Law, Morality and Autonomous Reasoning*. Hart Publishing.
- Sierra, M. and G.E. Berrios (1997) ‘Depersonalization: a conceptual history’ *History of Psychiatry*, 8:30, 213–229.
- Siewert, Charles P. (1998) *The Significance of Consciousness*. Princeton University Press.
- Simmons, A. John (2010) ‘Ideal and Nonideal Theory’ *Philosophy & Public Affairs*, 38:1, 5–36.
- Sinnott-Armstrong, Walter, ed. (2014) *Moral Psychology. Volume 4. Free Will and Moral Responsibility*. The MIT Press.
- Sizoo, Edith, ed. (2008) *Responsabilité et cultures du monde. Dialogue autour d’un défi collectif*. Éditions Charles Léopold Mayer.
- Skinner, Quentin (2002) ‘A Third Concept of Liberty’ *Proceedings of the British Academy*, 117, 2001 Lectures. Oxford University Press, 237-268.
- Slors, Marc (1996) ‘Why Dennett Cannot Explain What It Is To Adopt the Intentional Stance’ *The Philosophical Quarterly* (1950–), 46:182, 93–98.
- (2015) ‘Interpretivism and the meaning of mental state ascriptions: Comments on Bruno Mölder’s Mind Ascribed’ *Studia Philosophica Estonica*, 1–10.
- Smiley, Marion (1992) *Moral Responsibility and the Boundaries of Community. Power and Accountability from a Pragmatic Point of View*. The University of Chicago Press.
- Smith, Nicholas H., ed. (2002) *Reading McDowell. On Mind and World*. Routledge.

- Smith, Ken; Sandra Moriarty; Gretchen Barbatsis and Keith Kenney, eds. (2005) *Handbook of Visual Communication. Theory, Methods, and Media*. Lawrence Erlbaum Associates, Publishers.
- Smithies, Declan and Daniel Stoljar (2012a) 'Introspection and Consciousness: An Overview' in Smithies and Stoljar (2012b), 3–26.
- eds. (2012b) *Introspection and Consciousness*. Oxford University Press.
- Sneddon, Andrew (2006) *Action and Responsibility*. Springer.
- (2013) *Autonomy*. Bloomsbury.
- Sniderman, Paul M.; Piazza, Thomas; Tetlock, Philip E. and Kendrick, Ann (1991) 'The New Racism' *American Journal of Political Science*, 35:2, 423–447.
- Snyder, Jon R. (2009) *Dissimulation and the Culture of Secrecy in Early Modern Europe*. University of California Press.
- Sparrow, Tom and Hutchinson, Adam, eds. (2013) *A History of Habit. From Aristotle to Bourdieu*. Lexington Books.
- Spiegel, Herbert (1981) 'Hypnosis: Myth and Reality' *Psychiatric Annals*, 11:9, 16–23.
- Sripada, Chandra (2015) 'Moral Responsibility, Reasons, and the Self' in David Shoemaker, ed., *Oxford Studies in Agency and Responsibility*, Volume 3, 242–264. Oxford University Press.
- Stagl, Jakob Fortunat (2012) 'Die Personwerdung des Menschen: Anfänge im Römischen Recht' in Hans Thomas and Johannes Hattler, eds., *Personen: Zum Miteinander einmaliger Freiheitswesen*. Ontos Verlag, 89–109.
- Stanley, Damian; Elizabeth Phelps and Mahzarin Banaji (2008) 'The Neural Basis of Implicit Attitudes'. *Current Directions in Psychological Science*, 17:2, The Interface between Neuroscience and Psychological Science, 164–170.
- Stein, Edward (1996) *Without Good Reason. The Rationality Debate in Philosophy and Cognitive Science*. Clarendon Press.
- Stelzner, Werner, ed. (1993), *Philosophie und Logik. Frege-Kolloquien Jena 1989/1991*. Walter de Gruyter.
- Steup, Matthias (2000) 'Doxastic Voluntarism and Epistemic Deontology'. *Acta Analytica*, 15(24), 25–56.
- Strack, Friedrich, ed. (1994) *Evolution des Geistes: Jena um 1800. Natur und Kunst, Philosophie und Wissenschaft im Spannungsfeld der Geschichte*. Klett-Cotta.
- Strawson, P.F. (1958) 'Persons' in Gustafson (1964), 377–403.
- (1959) *Individuals. An Essay in Descriptive Metaphysics*. Routledge.
- (1962) 'Freedom and Resentment' *Proceedings of the British Academy*, xlviii, 187–211.
- Strawson, Galen (1993) 'The Impossibility of Moral Responsibility' *Philosophical Studies*, 75, 5–24.
- Stroebe, Wolfgang and Insko, Chester A. (1989) 'Stereotype, Prejudice, and Discrimination: Changing Conceptions in Theory and Research' in Bar-Tal,

- Graumann, Kruglanski and Stroebe (1989), Chapter 1, 3–34.
- Stroud, Barry (1996) ‘The Charm of Naturalism’ *Proceedings and Addresses of the American Philosophical Association*, 70:2, 43 – 55.
- Sturma, Dieter, ed. (2001) *Person. Philosophiegeschichte – Theoretische Philosophie – Praktische Philosophie*. Mentis.
- Sullivan, Shannon (2004) ‘From the Foreign to the Familiar: Confronting Dewey Confronting Racial Prejudice’ *The Journal of Speculative Philosophy*, New Series, 18:3, 193–202.
- (2006) *Revealing Whiteness. The Unconscious Habits of Racial Privilege*. Indiana University Press.
- Sullivan, Shannon and Schmidt, Dennis J., eds. (2008) *Difficulties of Ethical Life*. Fordham University Press.
- Sullivan, Shannon and Tuana, Nancy, eds. (2007) *Race and Epistemologies of Ignorance*. State University of New York Press.
- Sunstein, Cass R. (2015) ‘Fifty Shades of Manipulation’ *Journal of Marketing Behavior*, 1, 213–244.
- Supreme Court (1896) *Plessy v. Ferguson*, 163 U.S. 537 (1896).
- (1954) *Brown v. Board of Education of Topeka, Kansas*, 347 U.S. 483 (1954).
- (1955) *Brown v. Board of Education of Topeka, Kansas*, 349 U.S. 249 (1955).
- (1971) *Griggs v. Duke Power*, 401 U.S. 431 (1971).
- Swaine, Lucas (2016) ‘The Origins of Autonomy’ *History of Political Thought*, XXXVII:2, 216–237.
- Swindal, James (2012) *Action and Existence. A Case For Agent Causation*. Palgrave Macmillan.
- Tajfel, Henri (1969) ‘Cognitive Aspects of Prejudice’ *Journal of Biosocial Science*, 1, Suppl. 1, 173–191.
- Talisse, Robert B. (2005) *Democracy after Liberalism. Pragmatism and deliberative politics*. Routledge.
- Talvitie, Vesa (2009) *Freudian Unconscious and Cognitive Neuroscience. From Unconscious Fantasies to Neural Algorithms*. Karnac.
- Tauber, Alfred I. (2005) ‘The reflexive project: reconstructing the moral agent’ *History of the Human Sciences*, 18:4, 49–75.
- Tausk, Victor (1919/ 2008) *Beeinflussungsapparate. Zur Psychoanalyse der Medien*. Semele Verlag. [English translation: ‘On the Origin of the “Influencing Machine” in Schizophrenia’ *Journal of Psychotherapy Practice and Research*, 1:2, 185–206, Spring 1992.]
- Taylor, Charles (1979) ‘Atomism’ in Taylor (1985a), Chapter 7, 187–210.
- (1983) ‘The concept of a person’ in Taylor (1985a), Chapter 4, 97–114.
- (1985a) *Human Agency and Language. Philosophical Papers 1*. Cambridge University Press.

- (1985b) *Philosophy and the Human Sciences. Philosophical Papers 2*. Cambridge University Press.
- (1989) *Sources of the Self. The Making of the Modern Identity*. Cambridge University Press.
- (2004) *Modern Social Imaginaries*. Duke University Press.
- Taylor, James Stacey, ed. (2005a) *Personal Autonomy. New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*. Cambridge University Press.
- (2005) 'Introduction' in James Stacey Taylor (2005b), 1–29.
- Taylor, Richard (1963) *Metaphysics*.
- (2006) 'Determinism, a historical survey' in Borchert (2006), 4–23.
- Teo, Thomas (2005) *The Critique of Psychology. From Kant to Postcolonial Theory*. Springer.
- Terbeck, Sylvia (2017) 'Prejudice: Can We Cure It?' TedxTalk Plymouth University. 20 July 2019 <<https://www.tedxplymouthuniversity.com/sylvia-terbeck/>>.
- Terbeck, S.; Kahane, G.; McTavish, S.; Savulescu, J.; Cowen, P.J. and Hewstone, M. (2012) 'Propranolol reduces implicit negative racial bias' *Psychopharmacology*, 222, 419–24.
- Thalberg, Irving (1978) 'Hierarchical Analyses of Unfree Action' *Canadian Journal of Philosophy*, 8:2, 211–226.
- (1979) 'Socialization and Autonomous Behavior' *Tulane Studies in Philosophy*, 28, 21–37.
- Thaler, Richard H. and Sunstein, Cass R. (2008) *Nudge. Improving Decisions About health, wealth and happiness*. Penguin Books.
- Thiel, Udo (1997) 'Epistemologism' and early modern debates about individuation and identity' *British Journal for the History of Philosophy*, 5:2, 353–372.
- (2011) *The Early Modern Subject. Self-Consciousness and Personal Identity from Descartes to Hume*. Oxford University Press.
- Thiesmeyer, Lynn (2003a) 'Introduction: Silencing in discourse' in Thiesmeyer (2003b), 1–33.
- ed. (2003b) *Discourse and Silencing*. John Benjamins Publishing Company.
- Thomas, James and Brunnsma, David (2014) 'Oh, you're a racist? I've got a cure for that!' *Ethnic and Racial Studies*, 37:9, 1467–1485.
- Thomas, Yan (1998) 'Le sujet de droit, la personne et la nature. Sur la critique contemporaine du sujet de droit' *Le Débat*, 3:100, 85–107.
- Thomasson, Amie L. (2009) 'Non-descriptivism About Modality' in *The Baltic International Yearbook of Cognition, Logic and Communication*, 4, 200 Years of Analytical Philosophy, 1–26.
- Thompson, John (1990) *Ideology and Modern Culture*. Polity Press.
- Throop, C. Jason and Murphy, Keith M. (2002) 'Bourdieu and phenomenology. A critical assessment' *Anthropological Theory*, 2:2, 185–207.
- Thümmel, W. (2019) 'Mentalismus' in Ritter and Gründer (1971–2007), Volume 5,

- 1138.
- Tileagă, Cristian (2016) *The Nature of Prejudice: Society, Discrimination and Moral Exclusion*. Routledge.
- Tilly, Charles (1998) *Durable Inequality*. University of California Press.
- Timpe, Kevin; Griffith, Meghan and Levy, Neil, eds. (2017) *The Routledge Companion to Free Will*. Routledge.
- Todd, Patrick (2013) 'Manipulation' in LaFollette (2013) 3139–3145.
- Todd, William (1968) *Analytical Solipsism*. Martinus Nijhoff.
- Tobler, Christa (2008) *Limits and potential of the concept of indirect discrimination*. Luxembourg: Office for Official Publications of the European Communities.
- Tognazzini, Neal A. (2014) 'The Structure of a Manipulation Argument' *Ethics*, 124:2, 358–369.
- Tougas, Shelley (2011) *Birmingham 1963. How a photograph rallied civil rights support*. Compass Point Books.
- Toulmin, Stephen (1979) 'The Inwardness of Mental Life' *Critical Inquiry*, 6:1, 1–16.
- (1986) 'Self Psychology as a "Postmodern" Science' *Psychoanalytic Inquiry: A Topical Journal for Mental Health Professionals*, 6:3, 459–477.
- Trepagnier, Barbara (2010) *Silent Racism. How Well-Meaning White People Perpetuate the Racial Divide*. Second Edition. Paradigm Publishers.
- Tully, James (1999) 'The agonic freedom of citizens' *Economy and Society*, 28:2, 161–182.
- (2003) 'Wittgenstein and Political Philosophy: Understanding Practices of Critical Reflection' in Heyes (2003), Chapter 1, 17–42.
- Tuomela, Raimo (2002) *The Philosophy of Social Practices. A Collective Acceptance View*. Cambridge University Press.
- Tuomela, Raimo, and Kaarlo Miller (1988) 'We-intentions' *Philosophical Studies*, 53:3, 367–389.
- Turbanti, Giacomo (2017) *Robert Brandom's Normative Inferentialism*. John Benjamins Publishing Company.
- Turner, Patricia A. (1994) *Ceramic, Uncles and Celluloid Mammies: Black Images and Their Influence on Culture*. University of Virginia Press.
- Turner, Stephen P. (1998) 'Making Normative Soup out of Nonnormative Bones' in Turner (2002).
- (2002) *Brains/ Practices/ Relativism. Social Theory after Cognitive Science*. The University of Chicago Press.
- Twain, Mark (1884) *Adventures of Huckleberry Finn* in Twain, Mark (1982) *Mississippi Writings*. The Library of America,
- Tweney, Ryan D. (1997) 'Jonathan Edwards and determinism' *Journal of the History of the Behavioral Sciences*, 33:4, 365–380.
- Vanderschraaf, Peter and Sillari, Giacomo (2014) 'Common Knowledge' *The Stanford*

- Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/spr2014/entries/common-knowledge/>>.
- Velleman, J. David (1996) 'The Possibility of Practical Reason' *Ethics*, 106:4, 694 – 726.
- Van Houdt, Toon (2002) 'Word Histories, and Beyond: Towards a Conceptualization of Fraud and Deceit in Early Modern Times' in Toon van Houdt; Jan L. de Jong; Zoran Kwak; Marijke Spies and Marc van Vaeck, eds. (2002) *On the Edge of Truth and Honesty. Principles and Strategies of Fraud and Deceit in the Early Modern Period*, 1–32. Brill.
- Vera, Hernán and Joe R. Feagin, eds. (2007) *Handbook of the Sociology of Racial and Ethnic Relations*. Springer.
- Vico, Giambattista (1744) *Principj Di Scienza Nuova* in Vico (1953) *Opere*. Riccardo Ricciardi Editore.
- Völmicke, Elke (2005) *Das Unbewußte im Deutschen Idealismus*. Königshausen & Neumann.
- Vossenkuhl, Wilhelm (1974) *Wahrheit des Handelns. Untersuchungen zum Verhältnis von Wahrheit und Handeln*. Bouvier Verlag Herbert Grundmann.
- Wacquant, Loïc (2016) 'A concise genealogy and anatomy of habitus' *The Sociological Review*, 64, 64–72.
- Waibel, Violetta L. (2005) 'Unbewusstes oder Unaufrichtigkeit? Freud und Sartre im Widerstreit über Reflexivität und Präreflexivität des Bewusstseins' in Grundmann et al. (2005), 460–491.
- Waldron, Jeremy (2005) 'Moral Autonomy and Personal Autonomy' in Christman and Anderson (2005), 307–329.
- Walker, Margaret Urban (2007) *Moral Understandings. A feminist studies in ethics*. Second Edition. Oxford University Press.
- Wallace, R. Jay (1994) *Responsibility and the Moral Sentiments*. Harvard University Press.
- (2001) 'Normativity, Commitment, and Instrumental Reason', *Philosopher's Imprint*, 1:3, 1–26.
- Waller, Bruce N. (1990) *Freedom Without Responsibility*. Temple University Press.
- Walzer, Michael (1989) 'A Critique of Philosophical Conversation' in Walzer (2007), Chapter 2, 22–37.
- (2007) *Thinking Politically. Essays in Political Theory*. Yale University Press.
- Ware, Leland (2007) 'A Comparative Analysis of Unconscious and Institutional Discrimination in the United States and Britain' *Georgia Journal of International and Comparative Law*, 36:1, 89–157.
- Warnke, Georgia (1992) *Justice and Interpretation*. The MIT Press.
- Warren, Robert Penn (1965/ 1993) *Who Speaks for the Negro?* Yale University Press.
- Washington, Natalia and Daniel Kelly (2016) 'Who's Responsible for This? Moral Responsibility, Externalism, and Knowledge about Implicit Bias' in Brownstein and Saul (2016), Chapter 1.1, 11–36.

- Watkins, Eric (2005) *Kant and the Metaphysics of Causality*. Cambridge University Press.
- Watson, Gary, ed. (1982) *Free Will*. Oxford University Press.
- (1987) ‘Free Action and Free Will’ *Mind*, New Series, 96:382, 145–172.
- Watson, Martha (2004) ‘The Issue Is Justice: Martin Luther King Jr.’s Response to the Birmingham Clergy’ *Rhetoric & Public Affairs*, 7:1, 1–22.
- Wegner, Daniel M. (2002) *The Illusion of Conscious Will*. MIT Press.
- Wegner, Daniel M., & Wheatley, T. P. (1999) ‘Why it feels as if we’re doing things: Sources of the experience of will’ *American Psychologist*, 54, 480–492.
- Weinstein, D. (2007) *Utilitarianism and the New Liberalism*. Cambridge University Press.
- Welchman (1995) ‘Locke on Slavery and Inalienable Rights’.
- Wellman, David (2000) ‘From Evil to Illness: Medicalizing Racism’ *American Journal of Orthopsychiatry*, 70:1, 28–32.
- (2007) ‘Unconscious Racism, Social Cognition Theory, and the Legal Intent Doctrine: The Neuron Fires Next Time’ in Vera and Feagin (2007), Chapter 4, 39–65.
- Wempe, Ben (2004) *T.H. Green’s Theory of Positive Freedom. From Metaphysics to Political Theory*. Imprint Academic.
- Whitebook, Joel (2005) ‘Against Interiority: Foucault’s Struggle with Psychoanalysis’ in Gary Gutting (2005) *The Cambridge Companion to Foucault*. Second Edition. Cambridge University Press, Chapter 11, 312–347.
- Whyte, Lancelot Law (1960) *The Unconscious before Freud*. Basic Books.
- Widerker, David and Michael McKenna, eds. (2003) *Moral Responsibility and Alternative Possibilities. Essays on the Importance of Alternative Possibilities*. Ashgate.
- Wikforss, Åsa Maria (2001) ‘Semantic Normativity’ *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 102:2, 203–226.
- Williams, Garrath (2000) ‘Normatively Demanding Creatures: Hobbes, the Fall and Individual Responsibility’ *Res Publica*, 6, 301–319.
- Wilson, Mark (1982) ‘Predicate Meets Property’ *The Philosophical Review*, 91:4, 549–589.
- (2006) *Wandering Significance. An Essay on Conceptual Behavior*. Clarendon Press.
- Wilson, Robert A. and Keil, Frank C., eds. (1999) *The MIT Encyclopedia of the Cognitive Sciences*. The MIT Press.
- Wilson, Timothy D. (2002) *Strangers to Ourselves. Discovering the Adaptive Unconscious*. The Belknap Press of Harvard University Press.
- Wilson, Timothy D.; Samuel Lindsey and Tonya Y. Schooler (2000) ‘A Model of Dual Attitudes’ *Psychological Review*, 107:1, 101–126.
- Wisdom, John (1934) *Problems of Mind and Matter*. Cambridge University Press.
- Witt, Charlotte, ed. (2011) *Feminist Metaphysics. Explorations in the Ontology of Sex, Gender and the Self*. Springer.

- Wittenbrink, Bernd and Schwartz, Norbert, eds. (2007) *Implicit Measures of Attitudes*. The Guilford Press.
- Wittgenstein, Ludwig (1969) *On Certainty*. Blackwell Publishing.
- Wolf, Susan (1990) *Freedom Within Reason*. Oxford University Press.
- Wolff, Robert Paul (1970/ 1976) *In Defense of Anarchism*. Harper Torchbooks.
- Wolterstorff, Nicholas (1997) 'Obligations of Belief: Two Concepts' in Hahn (1997), 217–238.
- Wood, Allen W. (1991) 'Unsociable Sociability: The Anthropological Basis of Kantian Ethics' *Philosophical Topics*, Modern Philosophy, 19:1, 325–351.
- Woods, Tryon P. (2017) 'The Implicit Bias Theory of Implicit Bias Theory' *Drexel Law Review*, 10, 631–672.
- Woody, J. Melvin and Phillips, James (1995) 'Freud's "Project for a Scientific Psychology" After 100 Years: The Unconscious Mind in the Era of Cognitive Neuroscience' *Philosophy, Psychiatry, & Psychology*, 2:2, 123–134.
- Wright, Joshua D. and Douglas H. Ginsburg (2012) 'Behavioral Law and Economics: Its Origins, Fatal Flaws, and Implications for Liberty' *Northwestern University Law Review*, 106, 1033–1088.
- Wrightman, Lawrence S. (2008) 'Organized Psychology's Efforts to Influence the Supreme Court on Matters of Race and Education' in Adams et al. (2008), Chapter 2, 25–43.
- Wyer, Robert S. and Srull, Thomas K., eds. (1994/ 2014) *Handbook of Social Cognition. Volume 1: Basic Processes*. Second Edition. Psychology Press.
- Yanow, Dvora (1992) 'Silences in Public Policy Discourse: Policy and Organizational Myths' *Journal of Public Administration Research and Theory*, 2, 399–423.
- (2003) *Constructing 'Race' and 'Ethnicity' in America. Category-Making in Public Policy and Administration*. M.E. Sharpe.
- Yolton, John W. (1983) *Thinking Matter. Materialism in Eighteenth Century Britain*. Basil Blackwell.
- Young, Iris Marion (1990a) *Justice and the Politics of Difference*. Princeton University Press.
- (1990b) 'Moral Judgment and Unconscious Prejudice' *Social Philosophy Today*, 3, 297–310.
- (2000) *Inclusion and Democracy*. Oxford University Press.
- (2011) *Responsibility for Justice*. Oxford University Press.
- (2004) 'Responsibility and Global Labor Justice' *The Journal of Political Philosophy*, 12:4, 365–388.
- Young, Robert (1980a) 'Autonomy and the 'Inner Self'' *American Philosophical Quarterly*, 17:1, 35–43.
- (1980b) 'Autonomy and Socialization' *Mind*, 89:356, 565–576.
- (1986) *Personal Autonomy. Beyond Negative and Positive Liberty*. Croom Helm Ltd.

- Young-Bruehl, Elisabeth (1996) *The Anatomy of Prejudices*. Harvard University Press.
- Yousef, Nancy (2004) *Isolated Cases. The Anxieties of Autonomy in Enlightenment Philosophy and Romantic Literature*. Cornell University Press.
- Zagzebski, Linda Trinkaus (2012) *Epistemic Authority. A Theory of Trust, Authority, and Autonomy in Belief*. Oxford University Press.
- Zagorin, Perez (2009) *Hobbes and the Law of Nature*. Princeton University Press.
- Zahavi, Dan (2007) 'The Heidelberg School and the Limits of Reflection' in Heinämaa, Lähtenmäki and Remes (2007), 267–285.
- Zheng, Robin (2016) 'Attributability, Accountability, and Implicit Bias' in Brownstein and Saul (2016), Chapter 1.3, 62–89.
- Zimmerling, Ruth (2005) *Influence and Power. Variations on a Messy Theme*. Springer.
- Zimmerman, Michael J. (1997) 'Moral Responsibility and Ignorance' *Ethics*, 107:3, 410–426.

Summary

Confronting discrimination and the veil of prejudice

Modern liberal democracies hold dear the value of non-discrimination. But it may be difficult for them to confront practices of discrimination. One difficulty is that people may, *unknown* to them, perform practices of discrimination even if they truly believe that equality is an important ideal. In principle, this may affect everyone, as a citizen, as a representative of a company, or as someone performing a public role (a teacher, a civil servant). Psychological research indicates, for example, that employment recruiters may reject some applicants for job interviews on a discriminatory basis even if they do not realise this. I call this phenomenon the *veil of prejudice*, one that is hanging between what people know about their own practices – which does not include knowledge about their prejudices – and what they do *not* know about the *discriminatory* influences on their own practices: a veil that somehow prevents people from acquiring knowledge of the *hidden prejudices* that turn their practices into discriminatory ones. And therefore, the veil of prejudice indicates a boundary between agential self-knowledge and agential self-ignorance.

If the veil of prejudice makes sense, no attempt to deliberate with such people and persuade them is likely to reduce their discrimination practices. After all, these people are already committed to ideals of equality. There is no need to worry. Some scholars are already working on solutions: one of them would be administering an anti-prejudice pill. Such a solution is an example of what I call *non-deliberative actions*: they influence people having hidden prejudices in a way that escapes their ‘knowing’ (epistemic) abilities but succeeds in diminishing their discriminating behaviour. As I am interested in *non-deliberative* actions that might solve discrimination as a *moral* issue, I have explored the following question: What are the conditions for non-deliberative actions to count as holding responsible those people who perform practices of hidden prejudice? Identifying these conditions makes it possible to assess which actions are morally justifiable to confront discrimination and which are not, for example, to assess whether we may hold people responsible for their discrimination by administering an anti-prejudice pill to them.

In this thesis, I have argued, very briefly summarised, the following. I have examined what it could mean to hold people responsible for their hidden prejudices and what it could mean to hold them responsible in a non-deliberative way. I have explored these issues from two different theoretical perspectives: mentalism, as the standard perspective, and pragmatism, as a new perspective that I propose. From a mentalist perspective, knowing your prejudices is a binary matter: you know your prejudices, or you do not. And increasing your self-knowledge (for example, about your hidden prejudices) is an internal (i.e. reflective), factual and brain-based process. The starting view in mentalism is that people are not responsible for their hidden prejudices simply

because hidden prejudices belong to a special type of mental state to which people cannot have any reflective access. For that reason, people cannot be held responsible for them. However, the mentalist perspective fails in theorising self-knowledge because it ignores that having self-knowledge is deeply embedded within external (i.e. social), normative and practical processes. To put my criticism more concretely and somewhat simplified: a mentalist framework for theorising hidden prejudices induces philosophers and social psychologists to trace the origin of discriminatory practices to mental, more specifically *unconscious*, and ultimately neurological processes in the brain of *one* employment recruiter, thus ignoring that decisions on inviting applicants take place within a social context in which everyone knows the hidden, implicit norms that they communicate through ‘brief impersonal encounters, stares, vocal inflections, hostile laughter, or public reversals of private expectations’ (Ellison 1989:821). From a pragmatist perspective, self-knowledge is a social, normative and practical matter because it is embedded within the epistemic practices of a community. What does it mean to have prejudices? It depends on a community’s understanding of discriminative practices, the norms to which that community is committed and the extent to which that community is responsive to whatever challenges her understanding and norms. Knowing your prejudices is a matter of degree. Having hidden prejudices is only partly a matter of self-ignorance: it does not by itself preclude responsibility because communal practices offer all kinds of opportunities of getting to know one’s prejudices. It opens the possibility of holding people responsible for their hidden prejudices. I have explored how we might do this in a non-deliberative way. These explorations have been tentative, discussing a case study on how news photographs have influenced public opinion. But they indicate the potential for influencing people in a non-deliberative yet moral way. In short: to hold them responsible for their hidden prejudices.

Unravelling the veil of prejudice: mentalism and pragmatism

To answer my research question, I needed various key notions: not only ‘hidden prejudices’ and ‘non-deliberative actions’, which I already mentioned, but also ‘responsible agency’, ‘epistemic autonomy’, ‘rationality’ and ‘freedom’. I will briefly explain why I needed them.

First, my research question requires a moral criterion to assess the moral justifiability of any type of action to count as holding responsible. For this, I chose the notion of ‘responsible agency’: any type of action is *not* morally justifiable if it interferes with people’s responsible agency. As my example of administering an anti-prejudice pill suggests, a non-deliberative action might be problematic exactly because it influences people in a way that bypasses people’s epistemic abilities. Therefore, I decided to focus more narrowly on the *epistemic* dimension of responsible agency and carve out within this notion the more specific notion of ‘epistemic autonomy’. Next, I needed an analysis of what non-deliberative actions mean. Only then could I assess whether non-deliberative actions, perhaps by definition, interfere with epistemic autonomy and thus

with responsible agency.

Second, analysing ‘responsible agency’ and ‘epistemic autonomy’ was also required for a moral assessment of people’s responsibility for hidden prejudices. My initial description of the veil of prejudice seems to imply that people are not responsible for their hidden prejudice because hidden prejudices bypass people’s epistemic abilities. If this is correct, a non-deliberative action may still be performed as an act of confronting discrimination, not, however, as an act of holding them responsible for their hidden prejudices. So, as it turned out, both non-deliberative actions and hidden prejudices are conceptually connected because they both seem to interfere with epistemic autonomy.

I decided to explore these notions from two different frameworks: mentalism and pragmatism. As I understand a framework, this is not a specific theory but a conceptual architecture that functions as a deep background to a family of theories. The reason for discussing both frameworks is that mentalism presents the dominant but, in my view, deeply problematic framework for many contemporary views on responsible agency. In my view, pragmatism holds the promise of providing a better framework for views on responsible agency.

I have explored both frameworks at two different levels of analysis: a *metaphysical* and a *moral-political* level of analysis. My primary interest was with the moral-political level: to see how each framework explains to what extent either non-deliberative actions or hidden prejudice interferes, or not, with people’s epistemic autonomy. More specifically: to see how each framework would qualify, *in epistemic terms*, the influence that both non-deliberative actions and hidden prejudices exert on people’s actions; to see if this influence can somehow be qualified as part of people’s agential self-ignorance.

However, to better understand the differences between mentalism and pragmatism as explanatory frameworks at the moral-political level, more fundamental analysis was required at what I called the metaphysical level. At this level, I analysed their notion of epistemic autonomy as a notion that tells us what it means to act as *epistemically free* beings. I wanted to know whether their notion of epistemic autonomy is rooted in fundamentally different assumptions: on what it means *to have knowledge* (which I connected with the notion of ‘rationality’), *to act freely* (which I connected with the notion of ‘freedom’ in a metaphysical rather than a political sense), finally, to how rationality and freedom are related.

I briefly sketch both frameworks.

Mentalism is a framework for understanding responsible agency in terms of *internalist*, i.e. *intra*-subjective features of people’s mind/body. As a result, rationality is understood as the primary feature of an individual (brain). Mentalism is therefore committed to a methodological, perhaps even ontological, individualism. Its basic strategy in offering an *epistemic* explanation of the mind consists in analysing the mind in terms of mental states and qualifying these states as either conscious or unconscious. What distinguishes conscious states from unconscious states is that the first must be

described by including the epistemic first-person perspective of the knowing agent while the second can be described as mental but without referring to her epistemic first-person perspective. It means that beliefs are subject to a dichotomous categorisation: they are either conscious, and in that sense deliberate, or unconscious. It suggests that hidden prejudices would count as unconscious beliefs. This does not yet indicate the nature of the boundary between agential self-knowledge and agential self-ignorance; in other words, how difficult or how easy it is for the knowing agent to *acquire knowledge* about what she is self-ignorant about.

The pragmatist framework that I have presented is based on Robert Brandom's normative pragmatism. In this version, pragmatism understands responsible agency in terms of *externalist*, i.e. *inter*-subjective features. It is understood in terms of claim-making as a practical, normative and social phenomenon. First, claim-making is basically practical. Pragmatism distinguishes between knowing-that and knowing-how, i.e. between saying and doing or, in Brandom's vocabulary, between explicit and implicit meaning. From a pragmatist view, it is possible to make a claim by either *saying* or *doing* it. In the first case, the claim is *explicit* in what people say; in the second case, it is *implicit* in what people do. The traditional view is that claim-making in saying is more basic than claim-making by doing. Pragmatism reverses this order. Claims implicit in doing form the baseline for claims explicit in saying. Second, claim-making is normative: it is understood as an inferential phenomenon. To understand a claim means to understand how this claim functions in a *correct* inferential relation to other claims, which means: as either a premise or a conclusion. Third, claim-making is a social practice: it consists in mutual 'scorekeeping'. Members of a community treat each other as claim-makers, as beings that make certain claims of what are correct inferential relations between claims. Having claims, in a practical, normative and social sense, is not a static status. Whoever is born and raised in a community needs to master the claims of her community. And even adults continue to master claims. Being a claim-maker, therefore, is always a process of mastering new claims. One implication of this idea is that the attribution of a claim to a person is a contextual matter. What claims a person is exactly committed to, for example, whether she is prejudiced, depends on that person's context: her mastery of various concepts and the social context that allows her to master various concepts. It means that the attribution of claims (Brandom's technical term for beliefs), for example, that someone has hidden prejudices, cannot be subjected to any binary categorisation. In the case of hidden prejudices: what makes prejudices 'hidden' can have very different meanings.

In first exploring the mentalist perspective on my key notions, I started identifying, at the moral-political level of analysis, the standard *liberal* notion of responsibility. As I argued, it relies on a notion of autonomy (chapter 2): people are responsible for their actions which they performed *autonomously*. It is a *mentalist* notion of responsibility because it understands reflection as a psychological process, as introspection. Standard

notions of reflective autonomy usually provide a criterion for what counts as the proper source of autonomous actions, namely some version of a reflective will. Such notions, therefore, belong to the family of reflective volitional autonomy. As I am interested in epistemic autonomy, I suggested a model for reflective epistemic autonomy. It was defined in terms of conscious, deliberative introspection. It means that a person is epistemically responsible for her actions to the extent that she has (or could have had) *conscious* beliefs of whatever influences her motivations and actions. If this is not the case, then at least, for this reason, she cannot be said to be responsible for her actions. I also proposed that we understand non-deliberative actions as interfering with people's epistemic autonomy if they are manipulative. Next, I examined the conditions for non-deliberative actions to count as manipulative; and I proposed that this is the case if non-deliberative actions bypass people's normal capacities for deliberative, conscious reasoning.

Next, I explored the mentalist perspective on the veil of prejudice (chapter 3). My starting point was as follows. In recent decades, the research that a group of social psychologists has performed indicates the plausibility of the veil of prejudice: people perform practices of discrimination because they are influenced by what they call 'implicit biases', which is the social-psychological version of hidden prejudices. I have argued that a mentalist framework underpins the theory that informs this empirical research.

I then turned to the question of how implicit biases might also be understood as a social phenomenon. I discussed two different approaches: a mentalist-epistemic one (Fricker) and a pragmatist one (Bourdieu). While the notion of (hidden) prejudice is usually invoked to explain discriminatory practices, both approaches try to move beyond it and understand hidden prejudice as accounting also for practices of silencing (what Fricker calls 'epistemic injustice' and Bourdieu calls 'censorship'). In a critical discussion of both approaches, I made some suggestions for an improved notion of the veil of prejudice as a social phenomenon. I have not, however, attempted to integrate them as they start from basically different starting points. The epistemic-mentalist approach defines prejudices in internal terms, i.e. in terms of unconscious states, and as the product of psychological, equally unconscious processes. Bourdieu has explicitly criticised the mentalist approach (in sociology). Instead, his pragmatist approach understands prejudices in *external* terms, i.e. what people *do* and as the product of social processes.

Now that I had explained the veil of prejudice in mentalist terms, I wanted to put to the test the liberal-mentalist of responsibility, and more specifically, the notion of reflective epistemic autonomy (chapter 4): are people responsible for their implicit biases? Using the notion of reflective epistemic autonomy, the answer to this question seemed clear: people are not responsible for their implicit biases because they are unconscious. But this does not yet settle the issue. We need to solve yet another

question: whether people can *acquire knowledge* about their hidden prejudices. In other words: how porous is the boundary between agential self-knowledge and agential self-ignorance? Can people somehow access their unconscious mental states and turn them into conscious mental states? I argued that implicit biases provide an example of what might be called a *local* problem of agential self-ignorance. People can usually acquire knowledge about their agency through reflection and discover their motivations for what they previously did not know. Sometimes they cannot, for example, when they have implicit biases because these seem to belong to a class of unconscious mental states which people cannot become conscious of by definition. It poses a challenge to holding people responsible for their hidden prejudices.

Then I shifted my mentalist analysis towards the metaphysical level because I wanted to show that the notion of reflective epistemic autonomy is vulnerable to an even greater threat (still chapter 4). Discussing the standard debate on responsibility for implicit biases and its underlying mentalist assumptions led me to explore what poses a more radical challenge to the liberal-mentalist notion of responsibility: the *global* problem of agential self-ignorance, better known as the problem of determinism. It arises once liberal theorists accept a *naturalising* account: i.e. once they understand the conscious/unconscious distinction as an ontological (or, as I explored it here, *neurological*) distinction. The threat is best illustrated in the standard debate on free will and determinism, a debate in which thought experiments centring on manipulation provide an important argument against the claim that people have freedom and, therefore, responsibility: the Manipulation Argument. In the epistemic version of the problem of determinism, human rationality, which is understood as the process of acquiring beliefs, is explained in terms of two distinct epistemic processes: those that are fast, automatic and unconscious (System 1) and those that are slow, deliberative and conscious (System 2). In philosophical psychology, this has been called the dual-process or dual-system view. Epistemic autonomy consists of only the second type of process. The ambition of a naturalising account of human rationality is to explain, eventually, all conscious processes in terms of unconscious processes. Starting from this naturalising account of human rationality, I argued that the notion of reflective epistemic autonomy cannot explain how the knowing agent might overstep the boundary between her conscious mental states and her unconscious mental states.

After my discussion of mentalism at the metaphysical level of analysis, I shifted towards introducing and discussing pragmatism while remaining at the metaphysical level (chapter 5). Here I introduced Robert Brandom's normative pragmatism. To develop a criticism against the mentalist framework itself, I explained two ideas: that freedom is a two-level notion and that rationality is inferential. The first idea helped to open space for an alternative not yet considered option in the standard debate on determinism: performing free actions but not being responsible for them. The second idea of rationality as inferential was further spelt out in terms of modality. At the most

basic level, rationality is practical in a non-deliberative way ('meaning is use'). Even at this basic, practical level, rationality is inferential in the sense that people's actions display their grasp of certain concepts, of what can be inferred from what, what is a reason for what (the space of reasons). From a young age onwards, people learn to master many concepts because they *can* practice their concepts by trying, trying again and making mistakes (the Test-Operate-Test-Exit cycle). This implies that modality is a feature, not just of people's reasoning but of their practices and of the world itself. The idea of inferential rationality was then aligned with the idea of freedom as a two-level notion.

My aim was then to show how this alignment and the idea of people as knowers (in a normative-modal-practical sense) undermines the mentalist framework to the extent that this framework implies the determinism thesis (which I think it does). The basic problem for mentalism, as I understand it, is that it understands rationality and freedom as distinct issues. It implies, as I argue in my thesis in more detail, that mentalism cannot explain what it means that people are *knowers*. Instead, pragmatism understands rationality as intimately connected with what I called transcendental freedom. I argued that pragmatism, because of this intimate connection between rationality and freedom, can explain its own theoretical foundations, while mentalism cannot: mentalism can be used for a naturalist account of human reasoning, but this account cannot explain how this naturalist account itself is possible.

Moving back from the metaphysical to the moral-political level of analysis, I began to explore the pragmatist perspective on the various moral-political notions that are important to my research question: the veil of prejudice (chapter 6), epistemic autonomy (chapter 7) and non-deliberative actions (chapter 7).

For the pragmatist notion of the veil of prejudice, I picked up on two ideas from Brandom's normative pragmatics: first, that practices of rationality are intrinsically social practices and, second, that rationality is explicit in what we say and implicit in what we do (chapter 6). Each of these notions, by themselves, do not suffice to explain hidden prejudice. My strategy was to combine them in the sense that the veil of prejudice is explained in terms of implicit social practices. To avoid an explanation of 'implicit' in terms of 'unconscious' I then explored Bourdieu's meta-theoretical considerations for his notion of 'habitus'. I argued for an explanation of hidden prejudices as incorrect inferences based on commitments that people do not recognise as *their* commitments, not even as *commitments*. To explain hidden prejudice as a *social* phenomenon, I chose to reinterpret Brandom's idea of mutual scorekeeping. I argued that we should distinguish – next to scorekeeping as the attribution to others of commitments that are different from mine – a more basic kind of scorekeeping consisting in the attribution of collective commitments to others they share with me. The first type takes place against the background of the second. Both strategies combined resulted in a notion of the veil of prejudice whose description requires the inclusion of the epistemic perspective of the

knowing agent, not as an individual, atomist perspective, but as a social perspective.

Next, I have argued for a pragmatist notion of responsible agency, more specifically epistemic autonomy, in relation to hidden prejudices (chapter 7). Central to the notion of epistemic autonomy is inferential rather than introspective self-knowledge. People have self-knowledge, for example, because they understand the implications of their actions from various social perspectives. This notion of epistemic autonomy has several implications. First, ignorance is not the opposite of inferential self-knowledge but is already intrinsic: agential self-ignorance *is part of* people's agential self-knowledge. Second, the acquisition of knowledge is not an individual, atomistic, epistemic act but one which is already part of a process of socialisation. Third, these epistemic processes of socialisation do not absorb epistemic autonomy. Instead, the inferential nature of our knowledge allows the freedom to disagree. It also enables people to adopt a critical stance towards the socio-political order. Based on this notion of epistemic autonomy, it can be claimed that people are responsible, in various degrees, for their hidden prejudices.

Next, I considered whether, from a pragmatist perspective, holding people responsible for their hidden prejudices may consist in taking non-deliberative actions against them (still chapter 7). I argued for a notion of non-deliberative action that may influence people without manipulating them. Furthermore, I suggested a strategy for adapting this notion to socio-political contexts: it consisted in understanding *inferential rationality*, as what underpins non-deliberative actions, more specifically as *narrative rationality*, and therefore to understand non-deliberative action as narrative performance. To show how this strategy might work, I discussed a case study on news photographs in the civil rights movement and their non-deliberative impact on white audiences and their public opinion. I suggested that this case study illustrates how the pragmatist framework might support the idea of non-deliberative actions that are not manipulative, i.e. do not violate the norm of the freedom to disagree.

Mentalism and pragmatism: unravelling their differences

For my research question, mentalism and pragmatism provide two plausible but very different answers. The simplified mentalist answer is that people are not, and cannot be, responsible for their hidden prejudices because these influence people's beliefs and actions in a way that bypasses their conscious, deliberative abilities. Consequently, they cannot be held responsible. Here we get stuck in answering the research question. We do not arrive at the point at which we can answer the question 'under which conditions' because there is no 'holding responsible'.

The simplified pragmatist answer is that, based on a notion of inferential self-knowledge, people can be claimed to be responsible, in various degrees and in various forms, for their hidden prejudices. I proposed that non-deliberative actions may count as holding responsible if they do not violate people's autonomy as the freedom to disagree. To explain this type of autonomy, I specified that a certain kind of epistemic balance

should exist between those who perform the non-deliberative action and those whose epistemic autonomy is at stake.

Apart from the differences in assessing responsibility for hidden prejudice, differences between mentalism and pragmatism also arise in theorising non-deliberative actions. The administration of an anti-prejudice pill provides an example of what is, for mentalism, a typical non-deliberative action. The pill is effective by affecting people's brains, but without addressing people's conscious, deliberative abilities: and therefore, it interferes with their epistemic autonomy. For pragmatism, what I suggested might illustrate a non-deliberative action is the publication of news photographs. The impact of such photographs on audiences bypasses – from the *theoretical* perspective of mentalism – people's deliberative abilities. But from the theoretical perspective of pragmatism, it does not. Such photographs can have an impact precisely because people are claim-makers. To what extent people can make explicit the impact news photographs have on them is a different question. Difficulties they might have to express this impact does not imply that the news photographs manipulated them or bypassed their epistemic abilities.

Mentalism and pragmatism represent two different paradigms: depending on which paradigm you choose, it means seeing different things, such as what happens to people as epistemic beings when they are manipulated or when they perform discriminatory practices. Their differences run deep. They seem to resist a simple comparison and the rejection of one or the other framework. I nevertheless suggest that we reject mentalism and opt for pragmatism as a framework for unravelling the veil of prejudice and morally assessing actions to confront discrimination. Very briefly, I criticise the mentalist framework: it ignores prejudices as a social and normative phenomenon; it ignores the many intermediate forms of knowledge that people may have mastered of their discriminatory practices and, therefore, ignores the many degrees to which people can be held responsible for them; finally, it does not give conceptual room for understanding non-deliberative actions as holding responsible, whereas the pragmatist framework can. To conclude, pragmatism provides a framework that is more sensitive to epistemic, normative and social nuances. It can help analyse the complexities of hidden prejudices and the challenges of confronting discrimination.

Samenvatting in het Nederlands

Discriminatie bestrijden en de sluier van vooroordelen

Moderne liberale democratieën hechten veel waarde aan non-discriminatie. Maar het bestrijden van discriminerende praktijken kan moeilijk voor ze zijn. Eén moeilijkheid is dat mensen, zelfs als ze vinden dat gelijkheid een belangrijk ideaal is, discriminerend kunnen handelen *zonder dat ze het weten*. Dit kan in principe iedereen overkomen: als burger, als vertegenwoordiger van een bedrijf, of als iemand met een publieke functie (een leraar, een ambtenaar). Uit psychologisch onderzoek blijkt bijvoorbeeld dat wervingsfunctionarissen sollicitanten soms op discriminerende basis afwijzen en zich dit niet realiseren. Ik noem dit fenomeen de *sluier van vooroordelen*, een sluier die hangt tussen wat mensen weten over hun eigen handelingen – en dit omvat geen kennis over hun eigen vooroordelen – en wat ze *niet* weten over *discriminerende* invloeden op hun eigen praktijken: een sluier die op de een of andere manier voorkomt dat mensen kennis verwerven van de *verborgen vooroordelen* die hun praktijken tot discriminerende praktijken maken. De sluier van vooroordelen markeert daarom een grens tussen zelfkennis en zelfonwetendheid over het eigen handelen.

Mocht zo'n sluier van vooroordelen werkelijk bestaan, dan zal geen enkele poging om met zulke mensen te *delibereren*, hen met redenen te overtuigen, erin slagen hun discriminatiepraktijken te verminderen. Veel mensen met verborgen vooroordelen, lijkt het psychologische onderzoek te suggereren, zijn immers al overtuigd van gelijkheidsidealen. Maar sommige wetenschappers werken aan oplossingen: bijvoorbeeld het toedienen van een anti-vooroordeelpil. Een dergelijke oplossing is een voorbeeld van wat ik *niet-deliberatieve handelingen* noem: handelingen die mensen aanzetten tot aanpassing van hun gedrag door ze te beïnvloeden op een manier die aan hun zelfkennis ontsnapt. In *niet-deliberatieve handelingen* die discriminatie zouden kunnen oplossen, ben ik geïntereiseerd als een *morele* kwestie. Daarom heb ik de volgende vraag onderzocht: *Wat zijn de voorwaarden voor niet-deliberatieve handelingen om mensen verantwoordelijk te houden voor hun praktijken van verborgen vooroordelen?* Het expliciteren van deze voorwaarden maakt het mogelijk om te beoordelen welke handelingen moreel verantwoord zijn om discriminatie te bestrijden, en welke niet, om bijvoorbeeld te beoordelen of we mensen verantwoordelijk mogen houden voor hun discriminatie door het toedienen van een anti-vooroordeelpil.

Kort samengevat, heb ik in dit proefschrift het volgende betoogd. Ik onderzocht wat het kan betekenen om mensen verantwoordelijk te houden voor hun *verborgen vooroordelen* en om ze op een *niet-deliberatieve* manier verantwoordelijk te houden. Ik heb deze kwesties onderzocht vanuit twee verschillende theoretische perspectieven: mentalisme, als het standaardperspectief, en pragmatisme, als een nieuw perspectief dat ik voorstel. Vanuit een mentalistisch perspectief is het kennen van je vooroordelen een binaire zaak – of je kent je vooroordelen, of je kent ze niet – en het vergroten

van je zelfkennis (bijvoorbeeld over je verborgen vooroordelen) is een intern (d.w.z. reflectief), feitelijk en hersengerelateerd proces. De basisopvatting van het mentalisme is dat mensen niet verantwoordelijk zijn voor hun verborgen vooroordelen, simpelweg omdat verborgen vooroordelen tot een speciaal soort mentale toestanden behoren waartoe mensen geen reflectieve toegang hebben. Om die reden kunnen mensen er niet verantwoordelijk voor worden gehouden. Het mentalistische perspectief schiet echter te kort in het theoretiseren van zelfkennis: het miskent dat het hebben van zelfkennis diep verankerd is in externe (d.w.z. sociale), normatieve en praktische processen. Om mijn kritiek concreter en enigszins vereenvoudigd te verwoorden: een mentalistisch denkkader voor het theoretiseren van verborgen vooroordelen zet filosofen en sociaal psychologen ertoe aan de oorsprong van discriminerende praktijken te herleiden tot één wervingsfunctionaris, namelijk tot de mentale, meer specifiek onbewuste, en uiteindelijk neurologische processen in haar brein. Maar ze miskennen daarmee dat beslissingen over het uitnodigen van sollicitanten plaatsvinden in een sociale context waarin iedereen de verborgen, impliciete normen kent die worden gecommuniceerd door ‘korte onpersoonlijke ontmoetingen, starende blikken, stembuigingen, vijandig gelach of publieke omkeringen van persoonlijke verwachtingen’ (Ellison 1989:821). Vanuit een pragmatisch perspectief is zelfkennis een sociale, normatieve en praktische aangelegenheid, omdat ze ingebed is in de epistemische praktijken van een gemeenschap. Wat vooroordelen betekenen, hangt af van wat een gemeenschap begrijpt van discriminerende praktijken, van de normen waaraan die gemeenschap waarde hecht en van de mate waarin die gemeenschap reageert op alles wat haar begrip en normen uitdaagt. Het kennen van je vooroordelen is een kwestie van gradatie en het hebben van verborgen vooroordelen is slechts gedeeltelijk een kwestie van zelfonwetendheid: het sluit op zichzelf verantwoordelijkheid niet uit, want gemeenschappelijke praktijken bieden allerlei mogelijkheden om de eigen vooroordelen te leren kennen. Dit biedt de mogelijkheid om mensen verantwoordelijk te houden voor hun verborgen vooroordelen. Ik onderzocht hoe dit op een niet-deliberatieve manier kan worden gedaan. Dit onderzoek had grotendeels een verkennend karakter. Het bestond onder meer uit een kleine studie naar de manier waarop de invloed van nieuwsfoto’s op de publieke opinie kunnen worden begrepen als non-deliberatief. Hoezeer deze studie ook vragen openliet, het liet zien dat hier kansen liggen om mensen op een *niet-deliberatieve, maar morele* manier te beïnvloeden. Kortom: hen verantwoordelijk te houden voor hun verborgen vooroordelen.

De sluier van vooroordelen ontrafelen: mentalisme en pragmatisme

Om mijn onderzoeksvraag te beantwoorden had ik verschillende kernbegrippen nodig: niet alleen ‘verborgen vooroordelen’ en ‘niet-deliberatieve handelingen’, die ik al noemde, maar ook ‘verantwoordelijk handelingsvermogen’, ‘epistemische autonomie’, ‘rationaliteit’ en ‘vrijheid’. Ik zal kort toelichten waarom ik ze nodig had.

Ten eerste vergt mijn onderzoeksvraag een moreel criterium om te beoordelen wanneer een handeling moreel te rechtvaardigen is als een daad van ‘verantwoordelijk houden’. Hiertoe koos ik de notie van ‘verantwoordelijk handelingsvermogen’: een handeling is moreel niet te rechtvaardigen als zij dit verantwoordelijk handelingsvermogen ondermijnt. Zoals mijn voorbeeld van het toedienen van een anti-vooroordelpil suggereert, kan een niet-deliberatieve handeling problematisch zijn, juist omdat het mensen beïnvloedt op een manier die de zelfkennis van mensen omzeilt. Ik besloot daarom om te focussen op de *epistemische* dimensie van verantwoordelijke handelingsvermogen en binnen dit begrip de meer specifieke notie van ‘epistemische autonomie’ af te bakenen. Vervolgens had ik een analyse nodig van wat niet-deliberatieve acties betekenen. Alleen dan zou ik kunnen beoordelen of niet-deliberatieve handelingen, misschien per definitie, iemands epistemische autonomie en dus ook diens verantwoordelijk handelingsvermogen ondermijnen.

Ten tweede was het analyseren van ‘verantwoordelijk handelingsvermogen’ en ‘epistemische autonomie’ ook vereist om te beoordelen of mensen verantwoordelijk zijn voor hun verborgen vooroordelen. Mijn beschrijving van de sluier van vooroordelen lijkt te impliceren dat mensen niet verantwoordelijk zijn voor hun verborgen vooroordelen omdat verborgen vooroordelen ontsnappen aan de zelfkennis van mensen. Als dit juist is, kan een niet-deliberatieve actie nog steeds worden uitgevoerd als een daad om discriminatie te *bestrijden*, maar niet als een daad om hen *verantwoordelijk* te houden voor hun verborgen vooroordelen. Hiermee bleek dat zowel niet-deliberatieve handelingen als verborgen vooroordelen conceptueel verbonden zijn omdat ze allebei epistemische autonomie lijken te ondermijnen.

Ik besloot deze begrippen te onderzoeken vanuit twee verschillende denkkaders: mentalisme en pragmatisme. Zoals ik een ‘denkkader’ begrijp, is dit niet zozeer een specifieke theorie, maar een conceptuele architectuur die fungeert als een diepe achtergrond voor een familie van theorieën. De reden om beide denkkaders te bespreken is dat mentalisme het dominante, maar naar mijn mening zeer problematische denkkader vormt voor veel hedendaagse opvattingen over verantwoordelijk handelingsvermogen. In mijn visie belooft pragmatisme een beter denkkader te bieden voor opvattingen over verantwoordelijk handelingsvermogen.

Ik heb beide denkkaders op twee verschillende niveaus van analyse onderzocht: een metafysisch en een moreel-politiek niveau. Mijn eerste interesse ging uit naar het moreel-politieke analyseniveau: ik wilde zien hoe elk denkkader een verklaring kan bieden voor het eventueel ondermijnen van iemands epistemische autonomie door hetzij niet-deliberatieve handelingen, hetzij verborgen vooroordelen. Meer specifiek: ik wilde zien hoe elk denkkader een verklaring biedt, in epistemische termen, van de invloed die zowel niet-deliberatieve acties als verborgen vooroordelen uitoefenen op handelingen van mensen; ik wilde zien of deze invloed op de een of andere manier kan worden begrepen als onderdeel van de zelfonwetendheid van mensen over hun eigen handelen.

Om echter de verschillen tussen mentalisme en pragmatisme op moreel-

politiek niveau beter te begrijpen, was een meer fundamentele analyse nodig op wat ik het metafysische niveau heb genoemd. Op dit niveau analyseerde ik hun notie van epistemische autonomie als een notie die laat zien wat het betekent om te handelen als epistemisch vrije wezens. Ik wilde weten of hun notie van epistemische autonomie geworteld is in fundamenteel verschillende veronderstellingen over wat het betekent om kennis te hebben – een kwestie die ik verbond met de notie van ‘rationaliteit’ – en om vrij te handelen – een kwestie die ik verbond met de notie van ‘vrijheid’ in metafysische in plaats van politieke zin –, en tot slot over hoe rationaliteit en vrijheid met elkaar samenhangen.

Beide denkkaders schets ik kort.

Mentalisme is een denkkader om verantwoordelijk handelingsvermogen te begrijpen als internalistische, d.w.z. intra-subjectieve, kenmerken van de geest/het lichaam van mensen. Als gevolg hiervan wordt rationaliteit begrepen als het primaire kenmerk van een individu. Mentalisme accepteert daarom een methodologisch, misschien zelfs ontologisch, individualisme. Zijn basisstrategie om een epistemische verklaring van de menselijke geest te bieden, bestaat erin deze te analyseren in termen van mentale toestanden en deze toestanden te kwalificeren als bewust of onbewust. Wat bewuste toestanden onderscheidt van onbewuste toestanden, is dat de eerste moeten worden beschreven vanuit iemands eerstepersoonsperspectief, terwijl de tweede kan worden beschreven als mentaal, maar zonder te verwijzen naar haar eerstepersoonsperspectief. Het betekent dat overtuigingen binair kunnen worden ingedeeld: ze zijn ofwel bewust, en in die zin opzettelijk, ofwel onbewust. Dit zou betekenen dat verborgen vooroordelen als *onbewuste* overtuigingen gelden. Maar dit geeft nog niet aan wat de aard is van de grens tussen zelfkennis en zelfonwetendheid over het eigen handelen, met andere woorden, hoe moeilijk of hoe gemakkelijk het is voor een persoon om kennis te verwerven over iets waarover zij nog zelfonwetend is.

Het pragmatische raamwerk dat ik heb gepresenteerd, is gebaseerd op het normatieve pragmatisme van Robert Brandom. Pragmatisme in deze versie begrijpt verantwoordelijk handelingsvermogen als een *vermogen-om-te-beweren* en dit wordt op zijn beurt begrepen in termen van externalistische, d.w.z. intersubjectieve kenmerken, anders gezegd: als een praktisch, normatief en sociaal fenomeen. Ten eerste is het *vermogen-om-te-beweren* in wezen praktisch. Pragmatisme maakt onderscheid tussen weten-dat en weten-hoe, tussen ‘zeggen’ en ‘doen’ of, in Brandoms vocabulaire, tussen *expliciete* en *impliciete* betekenis. Vanuit een pragmatisch perspectief is het mogelijk om iets te beweren door het te *zeggen* of het te *doen*. In het eerste geval is de bewering expliciet in wat mensen zeggen, in het tweede geval impliciet in wat mensen doen. De traditionele opvatting is dat beweren door te *zeggen* fundamenteeler is dan beweren door te *doen*. Pragmatisme keert deze volgorde om. Beweringen die impliciet zijn in het doen, vormen de basis voor beweringen die expliciet worden gezegd. Ten tweede is beweren normatief: het wordt opgevat als een inferentieel fenomeen. Een bewering begrijpen

betekent begrijpen hoe deze bewering functioneert in een correcte inferentiële relatie tot andere beweringen, wat betekent: als een premisse of een conclusie. Ten derde is beweren een sociale praktijk: het bestaat uit wederzijds 'scores bijhouden'. Leden van een gemeenschap behandelen elkaar als 'beweerdere' ('claim-makers'), als wezens die een standpunt innemen over wat correcte inferentiële relaties tussen beweringen zijn. Een 'beweeder' zijn, in praktische, normatieve en sociale zin, is geen statische status. Wie is geboren en getogen in een gemeenschap, moet ingewijd worden in de overtuigingen (beweringen) van haar gemeenschap en deze leren beheersen. Dit is een kwestie van meesterschap ('mastery'). En zelfs volwassenen blijven oefenen in wat beweringen betekenen. Een 'beweeder' zijn is daarom altijd een proces om nieuwe beweringen onder de knie te krijgen. Dit betekent onder meer dat het toeschrijven van een bewering aan een persoon een contextuele aangelegenheid is. Welke bewering precies aan iemand kan worden toegeschreven, bijvoorbeeld of zij bevooroordeeld is, hangt af van de context van die persoon: van haar beheersing van verschillende concepten en van de sociale context die haar in staat stelt verschillende concepten te beheersen. Het betekent dat het toekennen van beweringen ('claims', dit is Brandoms technische term voor 'beliefs', overtuigingen), bijvoorbeeld van een bewering dat iemand verborgen vooroordelen heeft, niet binair gecategoriseerd kan worden. Wat vooroordelen 'verborgen' maakt, kan heel verschillende betekenissen hebben.

Bij het verkennen van het mentalistische perspectief op mijn kernbegrippen, begon ik met het identificeren, op het moreel-politieke niveau, van de *liberale* standaardnotie van verantwoordelijkheid die steunt op een notie van autonomie (hoofdstuk 2): mensen zijn verantwoordelijk voor hun handelingen die zij *autonoom* verrichten. Ik suggereerde dat het in wezen een *mentalistische* notie van verantwoordelijkheid is, omdat reflectie op psychologisch niveau wordt opgevat als introspectie. Standaardbegrippen van reflectieve autonomie stellen vaak dat het criterium voor een autonome handeling inhoudt dat zo'n handeling gebaseerd moet zijn in een reflectieve wil. Dergelijke begrippen behoren daarom tot de begrippenfamilie van *reflectieve wilsautonomie*. Omdat ik geïnteresseerd ben in epistemische autonomie, heb ik een model voorgesteld voor reflectieve *epistemische* autonomie. Dat werd toen gedefinieerd in termen van bewuste, deliberatieve introspectie. Het betekent dat een persoon epistemisch verantwoordelijk is voor haar acties in de mate dat ze bewuste overtuigingen heeft (of had kunnen hebben) over alles wat haar motivaties en acties beïnvloedt. Als dit niet het geval is, kan zij althans om deze reden niet verantwoordelijk worden gehouden voor haar daden. Ik stelde ook voor dat niet-deliberatieve handelingen te begrijpen als ondermijnend voor de epistemische autonomie van mensen als ze manipulatief zijn. Vervolgens heb ik onderzocht wat de voorwaarden zijn om niet-deliberatieve handelingen als manipulatief te beschouwen, en stelde ik voor dat dit het geval is als niet-deliberatieve handelingen de normale capaciteiten van mensen voor deliberatief, bewust redeneren omzeilen.

Vervolgens verkende ik het mentalistische perspectief op de sluier van vooroordelen

(hoofdstuk 3). Ik nam als uitgangspunt het onderzoek dat een groep sociaal psychologen de afgelopen decennia heeft verricht en dat de plausibiliteit van de sluier van vooroordelen aangeeft: mensen voeren discriminatiepraktijken uit omdat ze worden beïnvloed door wat zij ‘implicit biases’ noemen. Een ‘implicit bias’ is de sociaalpsychologische versie van wat ik in mijn betoog ‘hidden prejudice’ heb genoemd. Ik heb betoogd dat de theorie die ten grondslag ligt aan dit empirische onderzoek, is ingebed in een breder, mentalistisch denkkader.

Vervolgens ging ik in op de vraag hoe impliciete vooroordelen ook kunnen worden opgevat als een sociaal fenomeen. Ik besprak twee verschillende benaderingen: een mentalistisch-epistemische (Fricker) en een pragmatische (Bourdieu). Hoewel de notie van (verborgen) vooroordelen meestal wordt gebruikt om discriminerende praktijken te verklaren, proberen beide benaderingen verder te gaan en verborgen vooroordelen te begrijpen als een verklaring voor praktijken-van-verzwijgen (‘practices of silencing’). (Fricker spreekt over ‘epistemisch onrecht’ terwijl Bourdieu over ‘censuur’ spreekt.) In een kritische bespreking van beide benaderingen heb ik enkele suggesties gedaan voor een beter begrip van de sluier van vooroordelen als sociaal fenomeen. Ik heb echter niet geprobeerd ze te integreren, omdat ze uitgaan van fundamenteel verschillende uitgangspunten. De epistemisch-mentalistische benadering definieert vooroordelen in interne termen, d.w.z. in termen van onbewuste toestanden, en als het product van psychologische, evenzeer onbewuste processen. Bourdieu heeft expliciet kritiek geuit op de mentalistische benadering (in de sociologie). In plaats daarvan begrijpt zijn pragmatische benadering vooroordelen in externe termen, d.w.z. van wat mensen doen, en als het product van sociale processen.

Toen de sluier van vooroordelen eenmaal in mentalistische termen was uitgelegd, wilde ik de liberaal-mentalistic van verantwoordelijkheid op de proef stellen, en meer specifiek de notie van reflectieve epistemische autonomie (hoofdstuk 4): zijn mensen verantwoordelijk voor hun impliciete vooroordelen? Met behulp van de notie van reflectieve epistemische autonomie leek het antwoord op deze vraag duidelijk: mensen zijn niet verantwoordelijk voor hun impliciete vooroordelen omdat ze zich er niet van bewust zijn. Maar hiermee is het probleem nog niet opgelost. We moeten nog een andere vraag oplossen, namelijk of mensen *kennis kunnen verwerven* over hun eigen, verborgen vooroordelen. Met andere woorden: hoe doordringbaar is de grens tussen zelfkennis en zelfonwetendheid over het eigen handelen? Kunnen mensen op de een of andere manier toegang krijgen tot hun onbewuste mentale toestanden en ze veranderen in bewuste mentale toestanden? Ik betoogde dat impliciete vooroordelen een voorbeeld zijn van wat een *lokaal* probleem van zelfonwetendheid kan worden genoemd. Hoewel mensen over hun eigen handelen en hun eigen motivaties gewoonlijk door introspectie te weten kunnen komen wat zij daarvoor niet wisten, kunnen ze dat soms niet, bijvoorbeeld wanneer ze impliciete vooroordelen hebben, omdat deze tot een klasse van onbewuste mentale toestanden behoren waarvan mensen zich per definitie niet bewust kunnen

worden. Mensen kunnen dus niet verantwoordelijk worden gehouden voor hun verborgen vooroordelen.

Daarna verschoof ik mijn mentalistische analyse naar het metafysische niveau omdat ik wilde laten zien dat de notie van reflectieve epistemische autonomie kwetsbaar is voor een nog grotere bedreiging (nog steeds hoofdstuk 4). Toen ik het standaarddebat over verantwoordelijkheid voor verborgen vooroordelen ('implicit biases') en de onderliggende mentalistische veronderstellingen besprak, ben ik gaan onderzoeken wat een radicalere dreiging vormt voor het liberaal-mentalistische idee van verantwoordelijkheid: het *omvattende* (in onderscheid van: lokale) probleem van zelfonwetendheid, beter bekend als het probleem van determinisme. Dit ontstaat zodra liberale filosofen een naturalistische uitleg van het bewuste/onbewuste onderscheid accepteren als in wezen een ontologisch (of, zoals ik het hier heb onderzocht, neurologisch) onderscheid. Deze dreiging wordt het best geïllustreerd in het standaarddebat over vrije wil en determinisme, een debat waarin gedachte-experimenten rond manipulatie een belangrijk argument vormen tegen de bewering dat mensen vrijheid en dus verantwoordelijkheid hebben: het Manipulatieargument. In de epistemische versie van het probleem van determinisme wordt menselijke rationaliteit, uitgelegd als de processen van het verwerven van overtuigingen, begrepen in termen van twee verschillende epistemische processen: de processen die snel, automatisch en onbewust zijn (Systeem 1) en de processen die langzaam, deliberatief en bewust zijn (Systeem 2). In de filosofische psychologie wordt dit de opvatting van het duale-proces of het duale-systeem genoemd. Epistemische autonomie bestaat alleen uit het tweede type processen. De ambitie van een naturalistische beschrijving van de menselijke rationaliteit is om uiteindelijk alle bewuste processen te verklaren in termen van onbewuste processen. Ik betoogde dat, als we uitgaan van een naturalistische verklaring van de menselijke rationaliteit, het hele idee dat iemand door introspectie de grens tussen haar bewuste mentale toestanden en haar onbewuste mentale toestanden zou kunnen overschrijden, onderuitgehaald wordt en dat, kortom, het hele idee van reflectieve epistemische autonomie onmogelijk wordt.

Na mijn bespreking van mentalisme op het metafysische niveau van analyse, verschoof ik naar het introduceren en bespreken van pragmatisme, terwijl ik op het metafysische niveau bleef (hoofdstuk 5). Hier introduceerde ik het normatieve pragmatisme van Robert Brandom. Om kritiek op het mentalistische raamwerk te kunnen ontwikkelen, heb ik twee ideeën uitgelegd: dat vrijheid een begrip op twee niveaus ('two-level notion') is, en dat rationaliteit inferentieel is. Het eerste idee hielp om ruimte te maken voor een alternatieve, nog niet overwogen optie in het standaarddebat over determinisme: vrije handelingen verrichten, maar er niet verantwoordelijk voor zijn. Het tweede idee (rationaliteit als inferentieel) werd verder uitgewerkt in termen van modaliteit. Op het meest basale niveau is rationaliteit praktisch op een niet-deliberatieve manier (duiding is doen, 'meaning is use'). Zelfs op dit basale, praktische niveau is rationaliteit inferentieel in die zin dat de handelingen van mensen hun begrip van

bepaalde concepten laten zien: wat kan worden afgeleid uit wat, wat is een reden voor wat. Mensen leren van jongs af aan veel concepten te beheersen doordat ze hun concepten *kunnen* oefenen: door te proberen, opnieuw te proberen en fouten te maken (de ‘Test-Operate-Test-Exit’-cyclus). Dit impliceert dat modaliteit (in het geval van ‘beweerders’: *kunnen* oefenen) niet alleen een kenmerk is van beweringen of redeneringen, maar ook van menselijke praktijken en van de wereld zelf. Het idee van inferentiële rationaliteit verbond ik vervolgens met het idee van vrijheid als een notie op twee niveaus.

Mijn doel was toen om te laten zien hoe deze verbinding en het idee van mensen als ‘beweerders’ (in normatief-modaal-praktische zin) het mentalistische raamwerk ondermijnt voor zover dit denkkader ook het determinisme impliceert (en volgens mij is dat inderdaad het geval). Het fundamentele probleem van mentalisme, zoals ik het begrijp, is dat het rationaliteit en vrijheid als afzonderlijke kwesties ziet. Het impliceert, zoals ik in mijn proefschrift in meer detail betoog, dat mentalisme niet kan verklaren wat het betekent dat mensen ‘beweerders’ zijn. In plaats daarvan beschouwt pragmatisme rationaliteit als nauw verbonden met wat ik *transcendentale* vrijheid noemde. Ik betoogde dat pragmatisme, vanwege deze nauwe samenhang tussen rationaliteit en vrijheid, zijn eigen theoretische grondslagen kan verklaren terwijl mentalisme dat niet kan: mentalisme kan worden gebruikt voor een naturalistische verklaring van menselijke rationaliteit, maar deze verklaring kan niet verklaren hoe deze naturalistische verklaring zelf mogelijk is.

Daarna keerde ik van het metafysische terug naar het moreel-politieke analyseniveau en begon ik te verkennen wat het pragmatische perspectief is op de verschillende moreel-politieke noties die belangrijk zijn voor mijn onderzoeksvraag: de sluier van vooroordelen (hoofdstuk 6), epistemische autonomie (hoofdstuk 7) en niet-deliberatieve handelingen (hoofdstuk 7).

Voor de pragmatische notie van de sluier van vooroordelen heb ik twee ideeën uit Brandoms normatieve pragmatiek verder uitgewerkt: ten eerste dat rationaliteitspraktijken intrinsiek sociale praktijken zijn en ten tweede dat rationaliteit expliciet is in wat we zeggen en impliciet in wat we doen (hoofdstuk 6). Elk van deze begrippen is op zichzelf niet voldoende om verborgen vooroordelen te verklaren. Mijn strategie was om ze te combineren in die zin dat de sluier van vooroordelen wordt verklaard in termen van impliciete, sociale praktijken. Om een uitleg van ‘impliciet’ in termen van ‘onbewust’ te vermijden, verkende ik Bourdieu’s metatheoretische overwegingen voor zijn notie van ‘habitus’ en pleitte vervolgens voor een uitleg van verborgen vooroordelen als een praktijk van onjuiste gevolgtrekkingen op basis van overtuigingen (hier als vertaling van, in Brandom’s vocabulaire, ‘commitments’) die mensen niet erkennen als *hun* overtuigingen, en zelfs niet als overtuigingen. Om verborgen vooroordelen uit te leggen als een sociaal fenomeen, heb ik gekozen voor een herinterpretatie van Brandoms idee dat mensen elkaars overtuigingen bijhouden als een vorm van ‘scores bijhouden’. Ik betoogde dat we – naast ‘scores bijhouden’ als toeschrijving aan anderen van overtuigingen die

anders zijn dan de mijne – een meer basaal type van ‘scores bijhouden’ zouden moeten onderscheiden, een die bestaat uit het toeschrijven van collectieve overtuigingen aan anderen die zij met mij delen. Het eerste type speelt zich af tegen de achtergrond van het tweede. Beide strategieën gecombineerd resulteerden in een notie van de sluier van vooroordelen waarvan de beschrijving het epistemische perspectief van de ‘beweerder’ vereist, maar niet als een individueel, atomistisch perspectief, maar als een sociaal perspectief van de ‘beweerder’.

Vervolgens heb ik gepleit voor een pragmatische notie van verantwoordelijk handelingsvermogen in relatie tot verborgen vooroordelen (hoofdstuk 7). Deze begon bij een notie van epistemische autonomie waarin *inferentiële*, in plaats van *introspectieve*, zelfkennis centraal staat: mensen hebben zelfkennis in de mate dat ze de gevolgen van hun handelingen (kunnen) begrijpen, dat ze deze gevolgen vanuit andere perspectieven kunnen begrijpen en dat hun begrip impliciet of expliciet is. Deze notie van epistemische autonomie heeft verschillende implicaties. Ten eerste is zelfonwetendheid niet het tegendeel van inferentiële zelfkennis, maar is daarvan een inherent onderdeel: zelfonwetendheid over eigen handelen maakt deel uit van de zelfkennis over eigen handelen. Ten tweede is het verwerven van kennis geen individueel, atomistisch, epistemisch proces, maar een proces dat deel uitmaakt van een socialiseringsproces. Ten derde wordt epistemische autonomie niet volstrekt opgeslokt door deze epistemische processen van socialisering. In plaats daarvan biedt de inferentiële aard van onze kennis de vrijheid om het oneens te zijn. Het stelt mensen ook in staat een kritische houding aan te nemen ten opzichte van de sociaal-politieke orde. Op basis van deze notie van epistemische autonomie kan worden beweerd dat mensen in verschillende mate verantwoordelijk zijn voor hun verborgen vooroordelen.

Vervolgens overwoog ik of, vanuit een pragmatisch perspectief, niet-deliberatieve handelingen moreel te rechtvaardigen zijn als een manier om mensen verantwoordelijk te houden voor hun verborgen vooroordelen (nog steeds hoofdstuk 7). Ik pleitte voor een notie van niet-deliberatieve handeling die mensen kan beïnvloeden zonder ze te manipuleren. Vervolgens stelde ik een strategie voor om dit begrip aan te passen aan sociaal-politieke contexten: het bestond uit het begrijpen van inferentiële rationaliteit: als iets wat ten grondslag ligt aan niet-deliberatieve handelingen, meer specifiek als *narratieve rationaliteit*, en daarom om niet-deliberatieve handeling te begrijpen als narratieve handeling. Om te laten zien hoe deze strategie zou kunnen werken, gaf ik een beperkte analyse van enkele nieuwsfoto's in de burgerrechtenbeweging en hun niet-deliberatieve impact op het blanke publiek en hun publieke opinie. Ik suggereerde dat deze analyse illustreert hoe het pragmatische denkkader het idee zou kunnen ondersteunen van niet-deliberatieve acties die niet manipulatief zijn, d.w.z. niet in strijd zijn met de norm van de vrijheid om het oneens te zijn.

Mentalisme en pragmatisme: het ontrafelen van hun verschillen

Op mijn onderzoeksvraag leveren mentalisme en pragmatisme twee plausibele maar heel verschillende antwoorden op. Het vereenvoudigde mentalistische antwoord is dat mensen niet verantwoordelijk zijn, en niet kunnen zijn, voor hun verborgen vooroordelen omdat deze de overtuigingen en handelingen van mensen beïnvloeden op een manier die hun bewuste, deliberatieve vermogens omzeilt. Zij kunnen dan ook niet verantwoordelijk worden gehouden. Hier lopen we daarom vast bij het beantwoorden van de onderzoeksvraag. We komen niet toe aan het beantwoorden van onderzoeksvraag, omdat er geen sprake is van ‘verantwoordelijk zijn’ voor verborgen vooroordelen.

Het vereenvoudigde pragmatische antwoord, gebaseerd op een notie van inferentiële zelfkennis, is dat mensen in verschillende mate en in verschillende vormen verantwoordelijk zijn voor hun verborgen vooroordelen. Ik stelde voor dat een niet-deliberatieve handeling moreel gerechtvaardigd is als een daad van ‘verantwoordelijk houden’ als zij de autonomie van mensen, dus hun vrijheid om het oneens te zijn, niet schendt. Om dit soort autonomie uit te leggen heb ik aangegeven dat er een bepaald soort epistemische balans moet bestaan tussen degenen die de niet-deliberatieve handeling uitvoeren en degenen wiens epistemische autonomie op het spel staat.

Afgezien van de verschillen in het beoordelen van verantwoordelijkheid voor verborgen vooroordelen, doen zich ook verschillen tussen mentalisme en pragmatisme voor bij het theoretiseren van niet-deliberatieve handelingen. Het toedienen van een anti-vooroordeelpil is een voorbeeld van wat voor mentalisme een typische niet-deliberatieve handeling is. De pil is effectief dankzij het beïnvloeden van de hersenen, maar zonder de bewuste, deliberatieve vermogens van mensen te activeren: en daarom ondermijnt deze pil hun epistemische autonomie. Wat vanuit het pragmatisme zou kunnen illustreren wat een niet-deliberatieve handeling is, suggereerde ik, is de publicatie van nieuwsfoto's. De impact van dergelijke foto's op het publiek omzeilt – bezien vanuit het theoretische perspectief van mentalisme – het deliberatieve vermogen van mensen. Maar bezien vanuit het theoretische perspectief van pragmatisme is dit geen probleem. Zulke foto's kunnen impact hebben juist omdat mensen ‘beweerders’ zijn. In hoeverre mensen expliciet kunnen maken welke impact nieuwsfoto's op hen hebben, is een afzonderlijk vraagstuk. De moeilijkheden die ze zouden kunnen hebben om deze impact in woorden uit te leggen, betekent niet dat de nieuwsfoto's hen hebben gemanipuleerd of hun epistemische vermogens hebben omzeild.

Mentalisme en pragmatisme vertegenwoordigen twee verschillende paradigma's: afhankelijk van welk paradigma je kiest, zie je andere dingen, bijvoorbeeld in wat er met mensen als epistemische wezens gebeurt wanneer ze worden gemanipuleerd of wanneer ze discriminerende praktijken uitvoeren. De verschillen tussen beide denkkaders zijn diepgeworteld. Het lijkt daarom niet mogelijk om ze simpelweg te vergelijken en daarna het ene of het andere te verwerpen. Ik stel niettemin voor dat we het mentalisme afwijzen en kiezen voor het pragmatisme als denkkader voor het ontrafelen van de sluier

van vooroordelen en voor de morele beoordeling van handelingen in de bestrijding van discriminatie. Heel kort samengevat bekritiseer ik het mentalistische denkkader voor het miskennen van verborgen vooroordelen als een sociaal en normatief fenomeen, voor het miskennen van de vele tussenvormen van kennis die mensen zich hebben eigengemaakt van hun eigen discriminerende praktijken, en daarmee voor het miskennen van de vele gradaties waarin mensen verantwoordelijk kunnen worden gehouden, en ten slotte, voor het ontberen van conceptuele ruimte om niet-deliberatieve handelingen te begrijpen als daden van ‘verantwoordelijk houden’ terwijl het pragmatische denkkader dat wel kan. Kortom, pragmatisme biedt een epistemisch genuanceerder, een normatief gevoeliger en een sociaal realistischer denkkader waarmee we de complexiteit van verborgen vooroordelen en het verantwoord bestrijden van discriminatie kunnen analyseren.

Curriculum Vitae

Martin Jacobus Blaakman (1973) received his Master's degree in Philosophy from Radboud University, Nijmegen. He then worked for the Immigration and Naturalisation Service (1999–2007), mainly as a staff advisor, and after that for the National ombudsman of the Netherlands, primarily as a strategic (senior) advisor (2007–2019). In between, he worked for a brief period for the Ministry of the Interior and Kingdom Relations. Since 2019 he works as a lecturer for the University of Utrecht, teaching ethics and political philosophy for bachelor students and supervising students in the Applied Ethics Master for their theses and internships. For his thesis, he was supervised by prof. dr. Veit-Michael Bader, from the University of Amsterdam, and later by prof. dr. Bert van den Brink, from the University of Utrecht (based at University College Roosevelt), and dr. Joel Anderson, from the University of Utrecht.

Quaestiones Infinitae

**PUBLICATIONS OF THE DEPARTMENT OF PHILOSOPHY
AND RELIGIOUS STUDIES**

- VOLUME 21 D. VAN DALEN, *Torens en Fundamenten* (valedictory lecture), 1997.
- VOLUME 22 J.A. BERGSTRA, W.J. FOKKINK, W.M.T. MENNEN, S.F.M. VAN VLIJMEN, *Spoorweglogica via EURIS*, 1997.
- VOLUME 23 I.M. CROESE, *Simplicius on Continuous and Instantaneous Change* (dissertation), 1998.
- VOLUME 24 M.J. HOLLENBERG, *Logic and Bisimulation* (dissertation), 1998.
- VOLUME 25 C.H. LEIJENHORST, *Hobbes and the Aristotelians* (dissertation), 1998.
- VOLUME 26 S.F.M. VAN VLIJMEN, *Algebraic Specification in Action* (dissertation), 1998.
- VOLUME 27 M.F. VERWEIJ, *Preventive Medicine Between Obligation and Aspiration* (dissertation), 1998.
- VOLUME 28 J.A. BERGSTRA, S.F.M. VAN VLIJMEN, *Theoretische Software-Engineering: kenmerken, faseringen en classificaties*, 1998.
- VOLUME 29 A.G. WOUTERS, *Explanation Without A Cause* (dissertation), 1999.
- VOLUME 30 M.M.S.K. SIE, *Responsibility, Blameworthy Action & Normative Disagreements* (dissertation), 1999.
- VOLUME 31 M.S.P.R. VAN ATTEN, *Phenomenology of choice sequences* (dissertation), 1999.
- VOLUME 32 V.N. STEBLETSOVA, *Algebras, Relations and Geometries (an equational perspective)* (dissertation), 2000.
- VOLUME 33 A. VISSER, *Het Tekst Continuüm* (inaugural lecture), 2000.
- VOLUME 34 H. ISHIGURO, *Can we speak about what cannot be said?* (public lecture), 2000.
- VOLUME 35 W. HAAS, *Haltlosigkeit; Zwischen Sprache und Erfahrung* (dissertation), 2001.
- VOLUME 36 R. POLI, *ALWIS: Ontology for knowledge engineers* (dissertation), 2001.
- VOLUME 37 J. MANSFELD, *Platonische Briefschrijverij* (valedictory lecture), 2001.
- VOLUME 37A E.J. BOS, *The Correspondence between Descartes and Henricus Regius* (dissertation), 2002.
- VOLUME 38 M. VAN OTEGEM, *A Bibliography of the Works of Descartes (1637-1704)* (dissertation), 2002.
- VOLUME 39 B.E.K.J. GOOSSENS, *Edmund Husserl: Einleitung in die Philosophie: Vorlesungen 1922/23* (dissertation), 2003.
- VOLUME 40 H.J.M. BROEKHUIJSE, *Het einde van de sociaaldemocratie* (dissertation), 2002.

- VOLUME 41 P. RAVALLI, *Husserls Phänomenologie der Intersubjektivität in den Göttinger Jahren: Eine kritisch-historische Darstellung* (dissertation), 2003.
- VOLUME 42 B. ALMOND, *The Midas Touch: Ethics, Science and our Human Future* (inaugural lecture), 2003.
- VOLUME 43 M. DÜWELL, *Morele kennis: over de mogelijkheden van toegepaste ethiek* (inaugural lecture), 2003.
- VOLUME 44 R.D.A. HENDRIKS, *Metamathematics in Coq* (dissertation), 2003.
- VOLUME 45 TH. VERBEEK, E.J. BOS, J.M.M. VAN DE VEN, *The Correspondence of René Descartes: 1643*, 2003.
- VOLUME 46 J.J.C. KUIPER, *Ideas and Explorations: Brouwer's Road to Intuitionism* (dissertation), 2004.
- VOLUME 47 C.M. BEKKER, *Rechtvaardigheid, Onpartijdigheid, Gender en Sociale Diversiteit; Feministische filosofen over recht doen aan vrouwen en hun onderlinge verschillen* (dissertation), 2004.
- VOLUME 48 A.A. LONG, *Epictetus on understanding and managing emotions* (public lecture), 2004.
- VOLUME 49 J.J. JOOSTEN, *Interpretability formalized* (dissertation), 2004.
- VOLUME 50 J.G. SIJMONS, *Phänomenologie und Idealismus: Analyse der Struktur und Methode der Philosophie Rudolf Steiners* (dissertation), 2005.
- VOLUME 51 J.H. HOOGSTAD, *Time tracks* (dissertation), 2005.
- VOLUME 52 M.A. VAN DEN HOVEN, *A Claim for Reasonable Morality* (dissertation), 2006.
- VOLUME 53 C. VERMEULEN, *René Descartes, Specimina philosophiae: Introduction and Critical Edition* (dissertation), 2007.
- VOLUME 54 R.G. MILLIKAN, *Learning Language without having a theory of mind* (inaugural lecture), 2007.
- VOLUME 55 R.J.G. CLAASSEN, *The Market's Place in the Provision of Goods* (dissertation), 2008.
- VOLUME 56 H.J.S. BRUGGINK, *Equivalence of Reductions in Higher-Order Rewriting* (dissertation), 2008.
- VOLUME 57 A. KALIS, *Failures of agency* (dissertation), 2009.
- VOLUME 58 S. GRAUMANN, *Assistierte Freiheit* (dissertation), 2009.
- VOLUME 59 M. AALDERINK, *Philosophy, Scientific Knowledge, and Concept Formation in Geulincx and Descartes* (dissertation), 2010.
- VOLUME 60 I.M. CONRADIE, *Seneca in his cultural and literary context: Selected moral letters on the body* (dissertation), 2010.
- VOLUME 61 C. VAN SIJL, *Stoic Philosophy and the Exegesis of Myth* (dissertation), 2010.
- VOLUME 62 J.M.I.M. LEO, *The Logical Structure of Relations* (dissertation), 2010.
- VOLUME 63 M.S.A. VAN HOUTE, *Seneca's theology in its philosophical context* (dissertation), 2010.

- VOLUME 64 F.A. BAKKER, *Three Studies in Epicurean Cosmology* (dissertation), 2010.
- VOLUME 65 T. FOSSEN, *Political legitimacy and the pragmatic turn* (dissertation), 2011.
- VOLUME 66 T. VISAK, *Killing happy animals. Explorations in utilitarian ethics.* (dissertation), 2011.
- VOLUME 67 A. JOOSSE, *Why we need others: Platonic and Stoic models of friendship and self-understanding* (dissertation), 2011.
- VOLUME 68 N. M. NIJSINGH, *Expanding newborn screening programmes and strengthening informed consent* (dissertation), 2012.
- VOLUME 69 R. PEELS, *Believing Responsibly: Intellectual Obligations and Doxastic Excuses* (dissertation), 2012.
- VOLUME 70 S. LUTZ, *Criteria of Empirical Significance* (dissertation), 2012
- VOLUME 70A G.H. BOS, *Agential Self-consciousness, beyond conscious agency* (dissertation), 2013.
- VOLUME 71 F.E. KALDEWAIJ, *The animal in morality: Justifying duties to animals in Kantian moral philosophy* (dissertation), 2013.
- VOLUME 72 R.O. BUNING, *Henricus Reneri (1593-1639): Descartes' Quartermaster in Aristotelian Territory* (dissertation), 2013.
- VOLUME 73 I.S. LÖWISCH, *Genealogy Composition in Response to Trauma: Gender and Memory in 1 Chronicles 1-9 and the Documentary Film 'My Life Part 2'* (dissertation), 2013.
- VOLUME 74 A. EL KHAIRAT, *Contesting Boundaries: Satire in Contemporary Morocco* (dissertation), 2013.
- VOLUME 75 A. KROM, *Not to be sneezed at. On the possibility of justifying infectious disease control by appealing to a mid-level harm principle* (dissertation), 2014.
- VOLUME 76 Z. PALL, *Salafism in Lebanon: local and transnational resources* (dissertation), 2014.
- VOLUME 77 D. WAHID, *Nurturing the Salafi Manhaj: A Study of Salafi Pesantrens in Contemporary Indonesia* (dissertation), 2014.
- VOLUME 78 B.W.P VAN DEN BERG, *Speelruimte voor dialoog en verbeelding. Basisschoolleerlingen maken kennis met religieuze verhalen* (dissertation), 2014.
- VOLUME 79 J.T. BERGHUIJS, *New Spirituality and Social Engagement* (dissertation), 2014.
- VOLUME 80 A. WETTER, *Judging By Her. Reconfiguring Israel in Ruth, Esther and Judith* (dissertation), 2014.
- VOLUME 81 J.M. MULDER, *Conceptual Realism. The Structure of Metaphysical Thought* (dissertation), 2014.
- VOLUME 82 L.W.C. VAN LIT, *Eschatology and the World of Image in Suhrawardī and*

- His Commentators* (dissertation), 2014.
- VOLUME 83 P.L. LAMBERTZ, *Divisive matters. Aesthetic difference and authority in a Congolese spiritual movement from Japan* (dissertation), 2015.
- VOLUME 84 J.P. GOUDSMIT, *Intuitionistic Rules: Admissible Rules of Intermediate Logics* (dissertation), 2015.
- VOLUME 85 E.T. FEIKEMA, *Still not at Ease: Corruption and Conflict of Interest in Hybrid Political Orders* (dissertation), 2015.
- VOLUME 86 N. VAN MILTENBURG, *Freedom in Action* (dissertation), 2015.
- VOLUME 86A P. COPPENS, *Seeing God in This World and the Otherworld: Crossing Boundaries in Sufi Commentaries on the Qur'ān* (dissertation), 2015.
- VOLUME 87 D.H.J. JETHRO, *Aesthetics of Power: Heritage Formation and the Senses in Post-Apartheid South Africa* (dissertation), 2015.
- VOLUME 88 C.E. HARNACKE, *From Human Nature to Moral Judgement: Reframing Debates about Disability and Enhancement* (dissertation), 2015.
- VOLUME 89 X. WANG, *Human Rights and Internet Access: A Philosophical Investigation* (dissertation), 2016.
- VOLUME 90 R. VAN BROEKHOVEN, *De Bewakers Bewaakt: Journalistiek en leiderschap in een gemediatiseerde democratie* (dissertation), 2016.
- VOLUME 91 A. SCHLATMANN, *Shi'i Muslim youth in the Netherlands: Negotiating Shi'i fatwas and rituals in the Dutch context* (dissertation), 2016.
- VOLUME 92 M.L. VAN WIJNGAARDEN, *Schitterende getuigen. Nederlands luthers avondmaalsgerei als identiteitsdrager van een godsdienstige minderheid* (dissertation), 2016.
- VOLUME 93 S. COENRADIE, *Vicarious substitution in the literary work of Shūsaku Endō. On fools, animals, objects and doubles* (dissertation), 2016.
- VOLUME 94 J. RAJAJIAH, *Dalit humanization. A quest based on M.M. Thomas' theology of salvation and humanization* (dissertation), 2016.
- VOLUME 95 D.L.A. OMETTO, *Freedom & Self-Knowledge* (dissertation), 2016.
- VOLUME 96 Y. YALDIZ, *The Afterlife in Mind: Piety and Renunciatory Practice in the 2nd/8th- and early 3rd/9th-Century Books of Renunciation (Kutub al-Zuhd)* (dissertation), 2016.
- VOLUME 97 M.F. BYSKOV, *Between experts and locals. Towards an inclusive framework for a development agenda* (dissertation), 2016.
- VOLUME 98 A. RUMBERG, *Transitions toward a Semantics for Real Possibility* (dissertation), 2016.
- VOLUME 99 S. DE MAAGT, *Constructing Morality: Transcendental Arguments in Ethics* (dissertation), 2017.
- VOLUME 100 S. BINDER, *Total Atheism* (dissertation), 2017.
- VOLUME 101 T. GIESBERS, *The Wall or the Door: German Realism around 1800*, (dissertation), 2017.

- VOLUME 102 P. SPERBER, *Kantian Psychologism* (dissertation), 2017.
- VOLUME 103 J.M. HAMER, *Agential Pluralism: A Philosophy of Fundamental Rights* (dissertation), 2017.
- VOLUME 104 M. IBRAHIM, *Sensational Piety: Practices of Mediation in Christ Embassy and NASFAT* (dissertation), 2017.
- VOLUME 105 R.A.J. MEES, *Sustainable Action, Perspectives for Individuals, Institutions, and Humanity* (dissertation), 2017.
- VOLUME 106 A.A.J. POST, *The Journey of a Taymiyyan Sufi: Sufism Through the Eyes of Imād al-Dīn Aḥmad al-Wāsiṭī (d. 711/1311)* (dissertation), 2017.
- VOLUME 107 F.A. FOGUE KUATE, *Médias et coexistence entre Musulmans et Chrétiens au Nord-Cameroun: de la période coloniale Française au début du XXIème siècle* (dissertation), 2017.
- VOLUME 108 J. KROESBERGEN-KAMPS, *Speaking of Satan in Zambia. The persuasiveness of contemporary narratives about Satanism* (dissertation), 2018.
- VOLUME 109 F. TENG, *Moral Responsibilities to Future Generations. A Comparative Study on Human Rights Theory and Confucianism* (dissertation), 2018.
- VOLUME 110 H.W.A. DUIJF, *Let's Do It! Collective Responsibility, Joint Action, and Participation* (dissertation), 2018.
- VOLUME 111 R.A. CALVERT, *Pilgrims in the port. Migrant Christian communities in Rotterdam* (dissertation), 2018.
- VOLUME 112 W.P.J.L. VAN SAANE, *Protestant Mission Partnerships: The Concept of Partnership in the History of the Netherlands Missionary Council in the Twentieth Century* (dissertation), 2018.
- VOLUME 113 D.K. DÜRING, *Of Dragons and Owls. Rethinking Chinese and Western narratives of modernity* (dissertation), 2018.
- VOLUME 114 H. ARENTSHORST, *Perspectives on freedom. Normative and political views on the preconditions of a free democratic society* (dissertation), 2018.
- VOLUME 115 M.B.O.T. KLENK, *Survival of Defeat. Evolution, Moral Objectivity, and Undercutting* (dissertation), 2018.
- VOLUME 116 J.H. HOEKJEN, *Pars melior nostri. The Structure of Spinoza's Intellect* (dissertation), 2018.
- VOLUME 117 C.J. MUDDE, *Rouwen in de marge. De materiële rouwcultuur van de katholieke geloofsgemeenschap in vroege modern Nederland* (dissertation), 2018.
- VOLUME 118 K. GRIT, "Christians by Faith, Pakistani by Citizenship". *Negotiating Christian Identity in Pakistan* (dissertation), 2019.
- VOLUME 119 J.K.G. HOPSTER, *Moral Objectivity: Origins and Foundations* (dissertation), 2019.
- VOLUME 120 H. BEURMANJER, *Tango met God? Een theoretische verheldering van bibliodans als methode voor spirituele vorming* (dissertation), 2019.

- VOLUME I 21 M.C. GÖBEL, *Human Dignity as the Ground of Human Rights. A Study in Moral Philosophy and Legal Practice* (dissertation), 2019.
- VOLUME I 22 T. VAN 'T HOF, *Enigmatic Etchings. True Religion in Romeyn de Hooghe's Hieroglyphica* (dissertation), 2019.
- VOLUME I 23 M. DERKS, *Constructions of Homosexuality and Christian Religion in Contemporary Public Discourse in the Netherlands* (dissertation), 2019.
- VOLUME I 24 H. NIEBER, *Drinking the Written Qur'an. Healing with Kombe in Zanzibar Town* (dissertation), 2020.
- VOLUME I 25 B.A. KAMPHORST, *Autonomy-Respectful E-Coaching Systems: Fending Off Complacency* (dissertation), 2020.
- VOLUME I 26 R.W. VINKESTEIJN, *Philosophical Perspectives on Galen of Pergamum: Four Case-Studies on Human Nature and the Relation Between Body and Soul* (dissertation), 2020.
- VOLUME I 27 L.J. JOZIASSE, *Women's faith seeking life; Lived Christologies and the transformation of gender relations in two Kenyan churches* (dissertation), 2020.
- VOLUME I 28 M. KRAMM, *Balancing Tradition and Development. A deliberative procedure for the evaluation of cultural traditions in development contexts* (dissertation), 2020.
- VOLUME I 29 N. MYLES, *Communality, Individuality and Democracy: A Defense of Personism* (dissertation), 2020.
- VOLUME I 30 A. OEGEMA, *Negotiating Paternal Authority and Filial Agency: Fathers and Sons in Early Rabbinic Parables* (dissertation), 2021.
- VOLUME I 31 A.A. GOUDRIAAN, *'Seit ein Gespräch wir sind': Language and dialogical experience in Hegel* (dissertation), 2021.
- VOLUME I 32 E.H. MEINEMA, *Regulating Religious Coexistence. The Intricacies of 'Interfaith' Cooperation in Coastal Kenya* (dissertation), 2021.
- VOLUME I 33 K.D. TIMMER, *Thresholds and limits in theories of distributive justice* (dissertation), 2021.
- VOLUME I 34 M.J. BLAAKMAN, *Confronting Discrimination and Unravelling the Veil of Prejudice. The epistemic conditions of responsibility for hidden prejudices* (dissertation), 2021.

Quaestiones Infinitae

$\theta\pi$