Autonomy-Respectful E-Coaching Systems
Fending off Complacency

Bart A. Kamphorst

θπ

Autonomy-Respectful E-Coaching Systems

Fending off Complacency

Autonomie-respecterende e-coachingsystemen

(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 28 september 2020 des ochtends te 11.00 uur

door

Bart Anthony Kamphorst

geboren op 14 mei 1985
te Amersfoort

Promotor: Dr. J.H. Anderson

ISBN 978-94-6103-078-8

Contents

Preface **vii**

Introduction **1**

1 Self-Regulation Failure **9**
1.1 Self-Regulation Failure and Weakness of Will 10
1.2 Conceptualizing Self-Regulation Failure 17
1.2.1 The Discrepancy-Reducing Feedback Loop 26
1.2.2 Interventions 46
1.3 Reasons for Developing Interventions 49

2 E-Coaching Systems ***55***
2.1 Characterizing Coaching *57*
2.2 E-Coaching Systems in the Literature *58*
2.3 Defining E-Coaching Systems 63
2.4 The Socio-Technical Relationship 66

3 Ethical Concerns of E-Coaching **73**
3.1 Social Justice 74
3.2 Infringing Rights of Personal Autonomy 80
3.3 Unintended Diminishment of Personal Autonomy 88

4 Complacency and Self-Governance **93**
4.1 The Planning Theory of Agency 96
4.1.1 Synchronic Self-Governance 100
4.1.2 Diachronic Self-Governance 106
4.2 Heightened Risk of Complacency 108
4.2.1 Empirical Evidence 110
4.2.2 Complacency Defined 115

4.2.3 Automation-Related Decision Bias Revisited 118
4.3 Implications for Self-Governance 120
4.3.1 The Role of Vigilance in Self-Governance 121
4.3.2 Objection: Extended Self-Governance 123
4.4 Illustrative Examples 127
4.4.1 Neglecting to Assess Suggestions 129
4.4.2 Falling Short in Assessing Suggestions for Means-End Coherence 134
4.4.3 Falling Short in Assessing Suggestions for Consistency 139

5 Implications and Recommendations 147
5.1 Normative Implications 148
5.1.1 Implications for Agents 149
5.1.2 Implications for E-Coaching System Designers 151
5.1.3 Implications for Policymakers 153
5.2 Anti-Complacency Recommendations 157

Conclusion 171

References 177

Samenvatting 201

Acknowledgements 207

Curriculum Vitae 211

Quaestiones Infinitae 212

Preface

This dissertation was conducted as part of the interdisciplinary research project "Promoting Effective Intentions: Volitional Scaffolding, Implementation Intentions, and Bedtime Procrastination" (HLS grant #12013). That project, headed by Dr. Joel Anderson, was itself part of the overarching interdisciplinary research programme "Healthy Lifestyle Solutions," which was jointly funded by the Netherlands Initiative on Brain and Cognition (NWO) and Philips Research.

Most of the dissertation is new, unpublished work. However, some of the chapters do share commonalities with (co-authored) papers that have been published as journal articles, book chapters or conference/workshop proceedings. In particular, I wish to acknowledge the following.

Parts of the related work section in Chapter 2 find their origin in Kamphorst, Klein & Van Wissen (2014b), "Human Involvement in E-Coaching: Effects on Effectiveness, Perceived Influence and Trust", in H.S. Park, A.A. Salah, Y.J. Lee, L.-P. Morency, Y. Sheikh & R. Cucchiara (Eds.), *Human Behavior Understanding*, Volume 8749 of *Lecture Notes in Computer Science*, pp. 16—29, Springer International Publishing.

The argument in favor of a more narrowly construed definition of e-coaching systems formed the basis for Kamphorst (2017), "E-Coaching Systems: What They Are, and What They Aren't", published in in a special issue of Personal and Ubiquitous Computing on Supporting a Healthier Lifestyle with e-Coaching Systems, volume 21, issue 4, pp. 625–632.

A first exploration of the relation between personal autonomy and the use of e-coaching systems can be found in Kamphorst & Kalis (2015), "Why option generation matters for the design of autonomous e-coaching systems", in AI & SOCIETY, volume 30, issue 1, pp. 77–88.

Finally, elements from the subsection on privacy concerns in Chapter 3 were previously published in Anderson & Kamphorst (2014), "Ethics

of E-Coaching: Implications of Employing Pervasive Computing to Promote Healthy and Sustainable Lifestyles", in *Proceedings of the Third IEEE International Workshop on the Social Implications of Pervasive Computing for Sustainable Living*, pp. 351–356, IEEE Computer Society Press.

Introduction

Imagine we had technologies that were so intimately attuned to our bodily processes, behaviors and even mental states that they could *predict* with high levels of accuracy both *when* and *why* we are most likely to behave in conflict with our own standards, or, in other words, when and why we are most likely to exhibit *self-regulation failure*. Imagine further that these technologies could subsequently *intervene* in our lives in an attempt to steer us straight, down the path we had chosen for ourselves. Lastly, imagine that these technologies would interact with us, question us, inform us, and help shape our plans needed to attain our most desired goals.

To make the image more concrete, consider the following examples. Suppose we are unsure of how to proceed with implementing our life goal of being healthy. In this future world, we could ask technology for help and within moments receive a list of concrete suggestions for action, based on our prior choices, personality profiles and preferences. Or suppose we are hungry and tempted to grab a quick bite at a fast food restaurant. Instantly, an alarm goes off, reminding us of our policy to eat nutritious foods; a policy we have because we care about living a long and healthy life. Or consider having a hard time pulling away from aimlessly watching television late at night. Given that it has been a particularly long day that has left us feeling drained, the lights in the room automatically start dimming slowly, subtly changing into a reddish hue and ten minutes later the television displays a notification in the top right corner that if we start our bedtime routine now, we can still enjoy a good eight hours of sleep. Finally, consider needing a nudge to go running after dinner and being notified that one the neighbors is planning to go for a run as well, is looking for a running buddy, and whether we would like to join.

For some, this futuristic scenario cannot become reality fast enough;

for others, however, it sketches a picture of a world they dread. The fact of the matter, however, is that though at this juncture in time we are still quite far away from the more advanced scenarios, it is this view of the future that is driving much research in a variety of sectors, both in academia and in industry. The hope behind many of these research endeavors is that these technologies will help support individuals in their projects of self-improvement, such as living a healthy lifestyle or maintaining a sustainable household, which in turn is hoped to have positive effects on society, for example by lowering health care costs and reducing global warming. This is a gross oversimplification of course, but the reasoning behind many research proposals runs roughly along those lines.

As is the case with most if not all innovation, the development of these technologies is bound to bring individual and societal costs as well as benefits, and will raise a plethora of ethical concerns (e.g., about equal access, privacy, responsibility, but also about authenticity and personal autonomy) that deserve to be considered seriously. Luckily, the fact that these technologies for the most part are not yet reality, places philosophers of technology, ethicists and other theorists in a good position to make concrete contributions to ongoing practical debates. By taking a *prospective* perspective and thinking about and analyzing possible scenarios ahead of time, there is a real opportunity to engage in constructive discussions with designers, developers, software architects and engineers and to pro-actively contribute to shaping these technologies in ways that maximize the benefits, minimize the costs, and, importantly, mitigate the risks. The first step however is knowing what the discussions should be about. Which brings us to this dissertation.

Aims

In this dissertation, my overarching aim is to make a positive contribution to the "responsible innovation" of self-regulation support systems by (1) arguing that the continuous use of poorly designed self-regulation support systems could facilitate a form of *complacency* that reduces users' personal autonomy and (2) developing several principles to guide efforts to mitigate these risks. In pursuing this broad aim, I have four subordinate aims.

The first aim is to define and label the category of technologies I am considering in this work. Getting explicit about this at an early stage is important, as it helps to prevent muddled discussions. Moreover, having a clear picture of the different functional components that these technologies consist of will help us in later chapters to understand the ways in which these systems increase the risk of becoming complacent.

The second aim is to disentangle this particular concern about complacency from other ethical concerns that may arise with these technologies. In order to make progress on this front, I will make a distinction between three categories of concerns, viz. concerns about social justice, concerns about infringements of rights of autonomy, and, finally, concerns about the potential negative effects that the interplay between e-coaching technologies and their users may have on people's exercise of self-governance. This separation of concerns can help with structuring future discussions about the ethics of self-regulation support systems. Moreover, for the immediate purposes in this dissertation, having the broader context in place will help to bring out the features of the complacency concern more clearly.

The third aim is to provide a philosophical analysis of the ways in which self-regulation support systems run the risk of negatively affecting people's personal autonomy by inducing or facilitating complacency in people's practical reasoning. Central to this analysis will be the idea that people's capacity for making plans and future-directed intentions plays an integral part in our *cross-temporally extended agency*. As I will argue, this idea has certain normative implications for the responsibilities that agents have towards themselves to ensure that the intentions they form and the plans they make are in line with and because of the values they themselves hold. As a consequence, any technology that potentially contributes to people mistakenly abdicating these responsibilities ought to be scrutinized carefully.

Finally, the fourth aim is to present a number of concrete recommendations that can serve as inputs to the aforementioned ongoing discussions about the further development of self-regulation support technologies. These recommendations will not be the result of a "standard ethical analysis" (if such a thing exists at all) in which stakeholders are identified and different concerns are weighed and tested in a variety of ethical frameworks, but will rather be the synthesis of what is discussed in the earlier chapters. I will say more about the

normative status of these recommendations in the last section of this introduction, which is concerned with methodology.

What follows now is a brief outline of the chapters to come. After the outline, the last section of this introduction will focus on explicating a number of methodological and terminological choices I have made.

Outline

Chapter 1 begins with an explanation of why I choose to use the terminology of self-regulation failure over the philosophically more familiar notions of *akrasia* and *weakness of will*. Next, I elaborate on the concept of self-regulation failure, discuss its boundaries, and introduce a number of related terms. Then, with the terminology established, I discuss two common perspectives on why people think that self-regulation failure is a phenomenon that requires the development of (technological) interventions, and add a third perspective that is sometimes overlooked outside of philosophy. Finally, the chapter concludes by emphasizing that the aforementioned reasons in favor of developing interventions are only *pro tanto* reasons that may be outweighed by other, more weighty *pro tanto* reasons related to ethical concerns that arise with the development of said interventions.

In Chapter 2 I will reflect on the nature of the technologies that can play a supporting role in people's self-regulation. Here I will argue for making a distinction between what I will call "self-regulation facilitators" and "e-coaching systems". It is the latter category that is of most interest philosophically, but without the distinction in place, there is a risk that discussions about the ethics of support technologies will get muddled with examples from the former category. I therefore offer a new definition of e-coaching systems and provide an initial characterization of e-coaching systems in terms of their functional components. I will conclude the chapter by characterizing the socio-technical relationship between user and e-coaching system in light of the literature on the extended mind and the extended will.

With the definition of e-coaching systems in place, we then turn our attention the ethics of such systems. In Chapter 3, I provide a broad overview of different kinds of ethical concerns that may arise with the widespread adoption of e-coaching systems. While the concern

about complacency that is central to this dissertation is located on the level of individuals, the overview extends towards societal concerns as well. The purpose of this is to give a sketch of the ethical landscape, which will hopefully benefit future discussions on the ethics of support technologies in general.

As mentioned before, I will distinguish three categories of concerns. The first category of concerns relates to social justice, and includes concerns about equal access to support technologies. The second category of concerns relates to the potential of these technologies to infringe on rights of autonomy. Here I survey concerns about privacy as well as about coercion and manipulation. Finally, the third category of concerns pertains to the potential effects on people's exercise of self-governance stemming from the interplay between e-coaching technologies and their users. I end the chapter with a broad sketch of the complacency concern that will be central to Chapters 4 and 5.

In Chapter 4 I begin by introducing Bratman's Planning Theory of Agency as the model of self-governing agency that I will use for my analysis of the complacency concern. Once this conceptual framework is in place, I turn to elaborating the pair claims that a) the continuous use of poorly designed self-regulation support systems can facilitate complacency in people's practical reasoning, and b) that the resulting complacency undermines people's self-governance. Drawing on work from Jason Kawall (2006), the chapter also provides a more nuanced characterization of complacency. Finally, with the help of fictional cases, I proceed to identify different ways in which different functional components of e-coaching systems can contribute to the facilitation of self-governance-undermining complacency in people's practical reasoning.

Chapter 5 aims at synthesizing and further developing the lessons from the previous chapters. In particular, in the first part of the chapter I discuss the normative implications of the findings from Chapter 4 for users and designers of e-coaching systems, as well as for policymakers. Then, in the second part, I will make recommendations that aim to be concrete enough to serve both as input to ongoing discussions about the development of e-coaching technologies and as helpful guidelines in the development process itself. The recommendations, in brief, are concerned with ensuring ongoing consent, revealing the reasoning behind suggestions, increasing user awareness about system fallibility,

offering reassessment opportunities, and promoting reflection on suggestions.

Finally, in the Conclusion I recapitulate the main points from each chapter. I conclude with a number of final remarks about the contributions of this dissertation for the responsible innovation of e-coaching systems.

This concludes the outline. Before moving on to the first chapter, however, I would like to make a few points about the terminology that I will use throughout this dissertation and make explicit some of the methodological choices I have made.

Methodology, Normativity and Terminology

Let me start with a general point about methodology. As mentioned in the preface, this work has come about in an interdisciplinary context; a context spanning philosophy, computer science, and psychology. Though I suspect that this interdisciplinarity will resonate throughout this dissertation, both in content and sometimes style, I consider this dissertation foremost a philosophical endeavor. This means that though I will at times review empirical literature, I will not present any new empirical data here myself.[1] Nor will I uncover newly developed software applications.[2] Rather, my main methods are conceptual analysis (e.g., analyzing the relation between complacency and self-governance; see Chapter 4) and what has recently been called "conceptual engineering" (Burgess, Cappelen, and Plunkett (2019); e.g., assessing and improving the concept of "e-coaching systems"; see Chapter 2). My hope is that these efforts to elucidate will make a positive contribution to the debate about the ethics of e-coaching systems and related technologies.

A second point concerns the normativity of the recommendations I make in this dissertation. Making recommendations about a practice is saying something about how the world *should* be. Recognizing this is important for understanding the type of project we are engaged in, as is being explicit about the normative assumptions underlying

[1] But see for example Kamphorst, Klein, and Van Wissen (2014a); Kamphorst et al. (2014b); Kamphorst, Nauts, De Ridder, and Anderson (2018).

[2] But see, e.g., Procee, Kamphorst, Van Wissen, and Meyer (2014); Kamphorst, Nauts, and Blouin-Hudon (2017); Kamphorst, Anderson, and Dignum (prep).

the recommendations. So I want to be open and upfront about my assumption that personal autonomy is a normatively relevant concept. The concept itself has been conceptualized in many different ways, but what is shared almost across the board is that it is about people "leading their lives by their own lights" (Anderson, 2013, p. 1) and that it involves self-direction, self-constitution and the development of certain capacities such as practical reason and self-reflection. So, when I say that personal autonomy is normatively relevant, I am stating that I take our capacity for self-governance—or self-rule, self-determination, or however else one chooses to conceptualize personal autonomy—to be valuable in and of itself, to be worthy of respect, and to carry weight in a multitude of decision-making contexts. Furthermore, I take this to be a well-established position, despite the very many different articulations of it (compare for example Raz, 1986; Mele, 1995; Oshana, 2006; Christman, 2009). Of course, I do recognize that the universality and inviolability of personal autonomy can be (and is) contested, in both discussions about morality (e.g., MacIntyre, 1981; Raz, 1986) and policy-making (e.g., see Conly, 2013), but I will not concern myself with those discussions here. For all I am assuming is that people having a voice in how they live their lives has intrinsic value. And for this I will give no further argumentation in this dissertation.

What I will do further down the line is become more precise about what I take it to mean when we say that an agent is self-governing. In Chapter 4 I will elaborate on the normative pressures that govern agency. Doing so will lay bare the responsibilities that we, as human beings, have towards ourselves in order to meet the requirements for self-governance. In particular, I believe that it follows from valuing self-governance that we have a responsibility towards ourselves to critically assess suggestions for action that are made to us from sources outside of ourselves. We will return to this idea in Chapter 4, but I mention it here because it will be helpful to keep it in mind while reading the first few chapters as well.

Let me conclude with a few points on the terminology I use throughout this dissertation. Where it matters, and it matters in most places, I will be as precise as possible in my wordings. However, I would like to flag that in some instances, I will be relatively loose in my use of the verb "to know". A statement of the kind "The e-coaching system *knows* that its user has a plan to A" will not be uncommon. This use should

be understood colloquially and in a relatively weak sense, just so that there is a basic understanding on our part that the e-coaching system has some information about the user's state of mind.

Another point concerns the use of the word "agent," and this comment is directed mostly at readers familiar with the Artificial Intelligence literature. One of the more exciting aspects of interdisciplinary work is crossing over from one academic discipline to another and finding that other people are working on the same type of problem from a different perspective. On some occasions, it turns out that they are investigating the same phenomenon, but under a different heading. A good example of this is the examination by psychologists of cases of *akrasia* under the header of "the intention-behavior gap" (see Chapter 1). At other times, however, researchers from different disciplines will use the same labels for different concepts. An example of this is the word "validation," which can mean entirely different things in either a psychology or a data science context. And finally, there are those cases in which the same labels refer to concepts that share the same origin, but that do not, or no longer, have the same meaning. These cases are often the hardest to discover—and the differences between the conceptions often hard to decipher—as it will not be immediately obvious that these concepts are understood differently within each discipline. As it happens, however, it is this type of scenario that we often find at the intersection between philosophy and computer science. In both disciplines, for example, labels such as agency, agents, and autonomy play an important role. Yet, depending on whose work you read of course, the meaning of these concepts can diverge widely. With that said, I want to flag that I fully realize that support systems such as e-coaching systems can often be (and often are) conceptualized and described as "agents" (artificial, virtual or otherwise). In fact, I have used this terminology myself on several other occasions.[3] Here, however, I reserve the term predominantly for human actors. Whenever I refer to other types of agents, I will make that explicit with an adjective. Likewise, I reserve the specific terms personal autonomy and self-governance for human agents, but I will talk about autonomy in relation to technological systems.

With the preliminaries now out of the way, let us move on to the first chapter.

[3] For example, see Kamphorst et al. (2014a,b).

Chapter 1

Self-Regulation Failure

"Self-regulation failure is the major social pathology of our time."

Roy F. Baumeister and John Tierney

I began the Introduction by sketching the type of technologies with which this dissertation is concerned, namely computerized systems which will be intimately attuned to us in ways that support our self-regulation efforts. In this first chapter, my primary aim is to provide the context needed for understanding the growing interest in academia and industry alike towards developing these technologies, as well as the reasons individuals may have for engaging with such technologies. My secondary aim is to begin my argument that, despite *pro tanto* reasons in favor of these technologies, we also have reasons for critically reflecting on this development.

The outline of the chapter is as follows. In Section 1.1, I will explain why using the term *self-regulation failure* is preferred in this dissertation to the philosophically more familiar notion of weakness of will. In Section 1.2, I will provide more substance to the concept of self-regulation failure and introduce a number of related terms. Then, with the terminology established, I will move on to Section 1.3 to discuss the two most common perspectives on why people think that self-regulation failure is a phenomenon so problematic that it requires the development of (technological) interventions. Moreover, I will argue that also from a third perspective—one that is sometimes overlooked outside of philosophy—there are additional *pro tanto* reasons for thinking that interventions aimed at reducing self-regulation failure might be de-

sirable. Finally, I will begin my argument that despite these pro tanto reasons in favor of developing *self-regulation support systems*, there are also concerns to consider, specifically with a subset of these systems that meet a certain threshold level of sophistication and independence. In Chapter 2 I will elaborate on the characteristics of these *e-coaching systems*.

1.1 Self-Regulation Failure and Weakness of Will

In the Introduction, I characterized self-regulation failure as behaving in conflict with one's own standards. In the following section, I will unpack this characterization and shed more light on this theoretical construct. As I am borrowing the concept from psychology, however, it will be instructive to first show its contours by comparing it to two conceptions of *weakness of will* as found in the philosophical literature on action theory. This comparison will make it clear that self-regulation failure captures a wider variety of agentic failings than either conception of weakness of will and that it lays bare more of the complexity of the causal mechanisms underlying these failings. As the technological interventions that are under investigation in this dissertation aim to address a wide spectrum of failures via different intervention approaches, I will conclude that the language of self-regulation failure is to be preferred in this context to the philosophically more familiar terms.

To begin the section properly then, consider the following hypothetical: an agent has formed an intention to go to bed at 11 p.m. because she cares about getting enough sleep, but later finds herself in one of the following three scenarios.

- *Scenario 1:* Watching another episode of her favorite television series at her intended bedtime while being distinctly aware that doing so goes against her better judgment to go to bed presently.
- *Scenario 2:* Watching another episode of her favorite television series at her intended bedtime while convincing herself that she has earned it to watch this additional episode after having faced a difficult day at work, despite having resolved earlier that day

to stick to her intended bedtime and to resist the anticipated temptation of watching more television.

- *Scenario 3:* Watching another episode of her favorite television series, which started automatically without her noticing that it is currently her intended bedtime.

Case 1 exemplifies weakness of will as it has been traditionally understood, namely as *akrasia*: acting against one's own best judgment about what it is one should do. This conceptualization of weakness of will—I will refer to it as the *classical* account—was made famous particularly by Donald Davidson's seminal article "How is Weakness of the Will Possible?" and a version of it is still defended today by contemporary philosophers such as Alfred Mele (Mele, 2010, 2012). Roughly, on this account the concept of weakness of will is defined as an agent doing x intentionally rather than y, while both x and y are real possibilities open to the agent, and the agent judges that, all things considered, it would be better to do y than to do x. When applied to the real-world example of going to bed later than intended, we would say that the agent who at time t_1 (the agent's intended bedtime) judges it best, all things considered, to go to bed now rather than later, but does not go to bed while it was possible to do so, is acting *akratically*.

So, what is distinctive of the classical account is that it attributes weakness of will solely to agents who act in conflict with their *present judgment* about what is best to do *at the present moment*. In later times, however, philosophers such as Richard Holton (1999, 2009) and Alison McIntyre (2006) have recognized that conceptualizing weakness of will *as* akrasia overly narrows the concept's scope by focusing exclusively on the judgment in the present moment. The problem with this, they argue, is that the concept thereby fails to capture a range of instances of weakness of will where the rational defects "in the connections between deliberation, motivation, and action" (McIntyre, 2006, p. 284) are essentially *diachronic* in nature, in that agents who at time t_1 intend a certain course of action to take at time t_2, unreasonably steer themselves in a different direction when time t_2 arrives.

The cases that these authors have in mind are therefore ones where the failure of rationality lies not in an agent's present judgment about what is best to do, for that judgment may "have followed [the agent's] judgement of which outcome is most desirable" (Holton, 2009, p. 100).

Rather, the kind of failure they are after lies in succumbing to foreseen temptation by unreasonably revising a specific kind of intention that one formed in the past as a result of practical deliberation and in relation to the foreseen temptation. Put differently, they are concerned with cases in which agents display weakness of will not because they act akratically but because they fail to be *resolute* in the face of temptation.

In developing an alternative account of weakness of will, Holton builds on Michael Bratman's *planning theory of agency* (Bratman, 1987) and its central thesis that human agents are resource-limited beings who form future-directed plans and intentions to guide their actions over time (Bratman, 1987). On his view—following Stroud (2010) I will refer to this conception as the *revisionist* account—Holton spells out weakness of will in terms of "unreasonable revision of a contrary inclination defeating intention (a resolution) in response to the pressure of those very inclinations" (Holton, 2009, p. 78). In other words, being persuaded by a temptation that one had anticipated and had sought to guard oneself against by forming a resolution not to be persuaded by that very temptation.[1]

As an illustration of this kind of weakness of will, consider scenario 2. Here, there is no akrasia since watching another episode is aligned with the agent's current, updated judgment that watching the episode is, all things considered, the best action to take in the present moment. In the example, this is highlighted by the agent rationalizing her choice to delay her bedtime. On the revisionist account, however, the agent is exhibiting weakness of will because she over-readily revised her prior resolution "to go to bed at 11 p.m. and to resist the anticipated temptation of watching more television". As McIntyre points out, agents who display weakness of will in this way can actually be said to be worse off in some respect concerning motivation than their akratic counterpart (cf. the agent from scenario 1) because at least the akratic agent still lucidly holds in mind that what she is currently doing (e.g., watching more television) is not the best thing for her to do (McIntyre, 2006, p. 287).

Giving into temptation in this way, including the process of rationalizing one's shift in judgment, is a common behavioral pattern that can be found in many domains of life. Examples include giving in

[1]For Holton, a resolution is a second-order intention not to reconsider a first-order intention to ϕ.

to drinking alcohol despite having resolved to stay sober ("it's okay, we're celebrating!"), going to a party despite having resolved to study for exams ("relaxing now will help me study later"), or splurging on expensive sports equipment despite having resolved to save money for traveling ("I've worked hard and deserve to treat myself"). By shifting the focus from judgments to intentions, the revisionist account can accommodate scenarios like these in a way that the classical account cannot.

Moreover, the focus on intentions gives the revisionist account a way to explain how weakness of will can exist even when it is not possible to make a judgment about which course of action is best, such as with cases of indifference or incommensurability.[2] For even when two courses of action are equally good or equally poor, or when the courses of action are incomparable, it is still possible to intend one and not the other. Lastly, it lets the revisionist account include cases where the agent does not experience any inner conflict after rationalizing his or her behavior, but where weakness of will can be attributed anyway from a third-person perspective: "[w]e surely can ascribe weakness of will to a person who has vowed to give up smoking, and who blithely starts up again straight away, saying that they have changed their mind" (Holton, 2009, p. 83). This is in stark contrast to the akratic agent who always subjectively "experiences a disharmony between what he judges best, and what he does" (Kalis, 2011, p. 8). In this sense, the revisionist account has broadened the dimension of culpability that is central to weakness of will beyond the flagrant disregard of one's own judgment that is distinctive of cases of akrasia.

Still, like the classical account, the revisionist account understands culpability only within the scope of an agent failing to act in accordance with and because of a particular mental state, viz. a judgment or intention respectively. Yet there exist many other cases in which agents are culpable for a misalignment of their behavior with what they care about that is not to be analyzed in this way. Specifically, agents may be culpable for what Stroud (2010) calls "defects in plan design," such as not making an intention strong enough to resist temptation (resolution), not taking sufficient precautions to increase the likelihood of their following through on their intention, or for not making an intention at all.

[2] See Holton (2009, p. 79).

As an illustration, consider scenario 3. Here, the agent's failure clearly cannot count as an instance of akrasia, since she is not deliberately going against her current judgment. Moreover, since the agent has not formed a resolution to refrain from watching another episode, her acting does not fit the revisionist's definition of weakness of will. Yet the agent's behavior is not in conformity with her intention, and it is far from evident that she is entirely free from blame. Granted, she is not aware that her intended bedtime has passed, but this could also be taken as an indication of negligently failing to adequately monitor the time she spent watching television. Moreover, if we suppose that she knows herself well enough to know that she has a disposition to lose track of time, she can also be said to have failed in taking the proper precautions to prevent this scenario from occurring (e.g., to set an alarm).

As another example, consider a related case in which an agent again cares about getting enough sleep, but initially does not form an intention to go to bed at a specific time. Instead, in the morning, she merely intends to go to bed "on time" this evening, leaving it open for her to determine at a later stage of the day exactly how she wants to specify what "on time" should mean for her this evening. As Bratman has argued, this approach to planning can be perfectly rational for resource-limited beings such as ourselves.[3] Perhaps the agent in this scenario wants to get a sense of how she is feeling throughout the day, what her workload is like, or whether her friends are planning any social activities before she decides on a specific bedtime. She may well make up her mind at 10 p.m. to go to bed at 11 p.m., or even just as the clock strikes eleven, and still be rational. It is only when the agent, by her own lights, judges that *now* is the best time to make up her mind about her bedtime, but subsequently does not form a more specific intention, that she opens herself up to criticism. So, for instance, if the agent at 7. p.m. judges that she should make a more specific plan about when to go to bed but fails to make the appropriate specification at that time, she may be criticized for a defect in her planning.

As a final example, consider a variant of the case in which the agent knows herself well enough to know she might be tempted to spend the evening watching television mindlessly and that this behavior is likely to lead to diminished sleep. If at 7 p.m. she anticipates this temptation

[3]For example, see Bratman (1987, p. 29); Bratman (2007, p. 286).

but fails to make a resolution to guard herself from this temptation, she can be charged with a particular kind of rational failure. McIntyre, who recognizes that these cases exist, argues that scenarios like these may even constitute instances of akrasia: "If I see that a more forthright plan is necessary, I am deliberatively akratic when I fail to form the necessary resolution" (McIntyre, 2006, p. 298).

Scenarios such as scenario 3 and its variations are like scenario 2 in the sense that they emphasize the temporal dimension of human agency by highlighting the importance of reflecting on one's own dispositions in relation to anticipated future events and environmental contexts, and forming plans about how to deal with the anticipated scenarios. However, by drawing attention to the various ways in which planning defects other than unreasonably revising intentions can undermine self-direction, they lay bare more of the complexity involved in exerting one's cognitive and volitional capacities ahead of time in order to ward off temptation and successfully pursue goals. Because of this, studying the intricacies of cases such as scenario 3 is highly relevant for bettering our theoretical understanding of how goal-directed behavior can break down, which in turn is key to developing practical interventions aimed at helping people improve their goal-pursuit success by making more effective plans. In this regard, focusing exclusively on cases of weakness of will as understood by the classical or the revisionist account would be to ignore a range of cases and underlying mechanisms that are relevant to the subject matter of this dissertation.

Of course, one could choose to broaden the conception of weakness of will to include cases of defective plan design, but this might lead some philosophers to object on the grounds that doing so would dilute the conception to a point where weakness of will loses its distinctive character. Holton, for example, deliberately reserves the term weakness of will specifically for egregious cases in which people culpably succumb to the very temptation they had anticipated and sought to guard themselves against. He distinguishes these cases from cases of "mere caprice," an example of which would be a person who keeps switching between two television series, each time changing his mind about which one is the better show to watch. For cases of irresoluteness such as these, he thinks the charge of weakness of will is simply too strong (see Holton (2009, p. 77)). Similarly, along with McIntyre, he

could hold that failing to make a resolution is indeed a practical defect, just not weakness of will.

Without wishing to dismiss the relevance of a stricter definition for specific analytic purposes, engaging in a debate about which cases are "really worthy" to be labelled weakness of will is not conducive to my project, and it might not even contribute to our overall understanding of the phenomenon itself. In recent years, Mele and Holton have been going back and forth presenting various examples that they think should be included in an account of weakness of will, and have turned to *experimental philosophy* (e.g., Knobe, 2007; Sosa, 2007) in the hopes of settling the matter by examining what "the ordinary notion" of weakness of will consists in. Unfortunately, with Mele claiming that the collected data "indicate that this ordinary notion is considerably closer to a relatively standard conception of akrasia" (Mele, 2012, p. 398), but May and Holton countering that neither Mele nor Holton have provided accounts that fit the ordinary notion of weakness of will exactly (May and Holton, 2012), this debate currently has all the signs of becoming a terminological dispute.

Instead of engaging with the dispute, some other authors have sought to avoid it by radically moving away from the concept of weakness of will altogether. Neil Levy, for example, has argued that that weakness of will does not exist *as a psychological kind* because it "is simply a significant and salient manifestation of a much broader phenomenon, whereby rational agents come to act in ways that do not reflect the range and power of their rational processes" (Levy, 2011, p. 152). As such, he believes "the concept is useful neither for the explanatory purposes of psychology, nor for the practical purposes of increasing our ability to maintain self-control" (Levy, 2011, p. 148), and is best abandoned in favor of other, more empirically informed theories.

While it is unclear whether Levy is warranted in drawing his radical conclusion that the concept of weakness of will serves no purpose at all in either our scientific endeavors or in our self-understanding, I am sympathetic to his argument that there exist other theories of human behavior that are better able to capture the breadth and complexity of the rational failings that human beings exhibit. Especially for my purpose of giving an analysis of technological interventions that aim to support people in various ways with improving the alignment of their

behavior with what they care about, having in view the full extension of ways in which agents can fail to display well-aligned behavior is paramount. I therefore will make a move here paralleling Levy's and put aside the language of weakness of will in favor of the language of the existing theoretical construct of self-regulation failure that already encompasses the three scenarios introduced at the beginning of this section, as well as various other ways in which behavior can be misaligned that I will mention in the next section where I will provide more substance to the concept.

1.2 Conceptualizing Self-Regulation Failure

In the previous section, I advocated using the broader notion of self-regulation failure over the more narrow notion of weakness of will. The benefit, I argued, was to get into view the breadth and complexity of the agentic failings that human beings exhibit. My aim in this section is to outline the boundaries of the concept of self-regulation failure by analyzing a range of cases that fall inside and outside the scope of the concept. Given that I am borrowing the concept from psychology, the discussion will be guided by references to a number of key works from the psychological literature.

To understand what it is to fail at self-regulation, it is instructive to say more about how self-regulation is typically conceptualized. However, as different perspectives have originated in various subfields of psychology (e.g., educational, clinical, organizational, and health psychology [for discussion, see Boekaerts, Pintrich, and Zeidner (2000, ch.1)]), there are substantial variations in the definitions that are used in the literature, as illustrated by the following formulations.

- "Self-regulation refers to self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals" (Zimmerman, 2000, p. 14)
- "Self-regulation is the self's capacity for altering its behaviors" (Baumeister and Vohs, 2007, p. 115)
- "Self-regulation refers to the general process by which people adopt and manage various goals and standards for their thoughts, feelings, and behavior, and then ensure that these goals and standards are met" (Fujita, 2011, p. 352)

- "Self-regulation refers to the proximate motivational processes by which persons influence the direction, amount, and form of committed effort during task engagement" (Kanfer, 1990, p. 222)
- "Self-regulation refers to those processes, internal and/or transactional, that enable an individual to guide his/her goal-directed activities over time and across changing circumstances (contexts)" (Karoly, 1993, p. 25)

Together, these definitions certainly show the contours of the concept. Some definitions highlight that self-regulation involves change over time, others emphasize that self-regulation is of a cyclical nature, and others still stress that self-regulation works in service of the attainment of certain ends. However, the different emphases placed on the various aspects by these definitions, together with an ambiguity as to whether self-regulation is best understood as processes or as a capacity, make it difficult to grasp what the core components of self-regulation are. As Metcalfe and Mischel have said, self-regulation "is a topic in danger of losing its boundaries" (Metcalfe and Mischel, 1999, p. 4). To remedy this, it will be constructive to consider distinctive features that theorists generally agree upon as *constitutive features* of self-regulation, and to separate those from *empirically contingent features* that are relevant for self-regulation in human beings. The constitutive components can be characterized as follows:

> *Self-Regulation.* Self-regulation is a set of processes that constitute a discrepancy-reducing feedback loop that involves forms of *self-monitoring* and *self-direction*, and is aimed at the attainment of one's own ends over time.

From this process-based definition, we can subsequently derive "capacity talk" by understanding the capacity for self-regulation as having what is necessary for forming such a process-based loop and to be successful at least some of the time. A "well-developed" capacity can subsequently be expressed in terms of a "self-regulation success rate". At this point, however, I have to make a small detour from the main line of reasoning in order to clarify what I mean by self-regulation success. This is important, because how we understand success also affects the conceptualization of self-regulation failure.

One way of conceptualizing success would be to link it to the actual attainment of the end, so that whether or not one meets one's end

is the defining difference between success and failure. This position seems to follow for instance from Fujita's definition, which includes the notion that self-regulation *ensures* the attainment of the end. To see that there is some initial credence to this idea, consider the following example. Suppose someone adopts a goal to complete a marathon, and, at a later point in time, completes a marathon because of having that goal. In this instance, we can, correctly I believe, ascribe self-regulation success to this individual.

However, tying success closely to the attainment of ends gets one into trouble when one considers everything that happens between the moment an end is adopted and the moment it is either attained or abandoned. In what follows, I will explain the problems that arise with this conceptualization and offer a alternative that avoids these problems and aligns with some of the other definitions, particularly those that highlight the temporality of self-regulatory processes (e.g., Karoly's definition).

To begin, let us consider what it would mean for the conceptualization of self-regulation failure if self-regulation success were to be understood in absolute terms of attaining one's ends. Failure could then be defined as follows:

> *Self-Regulation Failure with Attainment-Based Criterion.* Self-regulation failure is not attaining one's self-chosen end as a result of a non-normally functioning or breakdown of one or more self-regulation processes, on the counterfactual condition that, had these processes functioned normally, the attainment of the end would have been secured.

At first glance, this characterization may seem apt, as it covers a subset of cases that involve failures of the kind that intuitively should be considered failures of self-regulation. For example, having to withdraw from a planned marathon run due to being ill-prepared as a result of non-normally functioning self-regulatory processes would be qualified as self-regulation failure on this conception, and this strikes me as right.

However, the characterization of self-regulation failure with the attainment criterion does not fit well with the temporal and iterative nature of the self-regulation feedback loop (see Section 1.2.1), as it does not provide a natural vocabulary to express problems that occur *during* the pursuit of one's end. For example, what do we say about

the person who ultimately does complete the marathon, but who, in a preparatory stage, finds himself going through a phase in which, to his own regret, he cannot motivate himself to put in the training hours he had scheduled for himself? Despite ultimately attaining his self-chosen end and therefore having success, this person surely is, by his own lights, failing in some regard to his self-regulation during that particular time of his life.

Problems also arise for positive cases of self-regulation when success and failure are conceived of in absolute terms of the attainment of ends. Consider for example two athletes, both of whom aim to qualify for a national tournament, but only one of whom considers the national tournament to "merely" be a necessary means towards her ambitious, long-term end of qualifying for the upcoming Olympic games. Suppose both athletes qualify for the national tournament, but that the more ambitious athlete, after another year of diligent training does not meet the qualifying standards for the Olympics. Should we say that the less ambitious athlete was more successful in her self-regulation because she attained her end, while the more ambitious athlete did not? Surely we want to be able to credit the more ambitious athlete for all the work she put in working towards the Olympics, even though she ultimately did not attain her end.

To avoid problems like the ones just discussed, a better approach to conceiving self-regulation success and failure is to place them on a continuum in relation to the discrepancy between input and reference value in the feedback loop, i.e. whether one is moving towards or away from one's end. This line of thinking actually follows quite naturally from the very notion of a self-regulation feedback loop, considering that reducing the discrepancy between input and a reference value is the aim of any negative feedback loop (I will say more about this in Section 1.2.1). The "self-regulation success rate" I mentioned earlier this section is then not a measure of how often ends are attained, but a measure of how effectively one can reduce the discrepancy or keep the discrepancy within a certain range over time.[4]

Thinking of success and failure in relation to the discrepancy of

[4]The stipulation about keeping the discrepancy within a certain range is there to accommodate ends that require maintenance of a discrepancy rather than a reduction, in the way that a thermostat's end is to keep the discrepancy between an observed temperature and the set temperature within a certain range. I will say more about this idea in relation to self-regulation below.

the feedback loop, self-regulation failure could then be characterized as follows:

> *Self-Regulation Failure with Progress-Based Criterion.* Self-regulation failure is any breakdown or non-normal functioning of self-regulatory processes that results in an increased discrepancy between one's behavior or mood or thought (which enter the feedback loop as subjective perceptions) and the reference value (derived from a relevant standard), on the counterfactual condition that, had said processes functioned normally, the discrepancy would have been smaller.

Seeing self-regulation failure in this light, as "intermittent" rather than absolute failure of the functioning of a system, makes it possible to neatly accommodate the two examples that were problematic before. The marathon runner who temporarily cannot motivate himself to put in the training hours is failing to self-regulate because by not training he is in effect moving away from attaining his self-chosen end. The ambitious athlete can be attributed self-regulation success for qualifying for the national tournament, but also for the work she puts in afterwards, because all those steps moved her closer to the attainment of her end, even though, ultimately, she did not attain it.

Besides accommodating the above-mentioned cases, the conception also has a number of other features that are worth discussing. First, notice that on this conception a increased discrepancy on its own does not automatically imply failure, because the counterfactual condition states that the discrepancy has to be causally "linked up" to a breakdown or non-normal functioning of relevant processes. This sets apart true failures of self-regulation from cases where the discrepancy increases due to external factors that are outside of the agent's control. For example, a family emergency might frustrate one's short-term plans to finish a scientific manuscript, but this does not qualify as self-regulation failure on this conception, and rightly so.

The conception also distinguishes cases of actual failure from cases in which an increased discrepancy is the unfortunately outcome of normal functioning of the self-regulatory processes. Consider, for instance, learning to throw free throws in basketball (Baumeister, 1984), which "requires careful regulation of one's balance and hand-eye coordination" (Fujita, 2011, p. 353). The objective of getting the ball into the hoop is clear, but the learning will likely not be a linear process

of getting closer and closer to making the shot with each attempt. Exploring different ways of attaining balance and manipulating the ball with one's throwing hand might inadvertently lead one to miss a free throw by a larger distance than the one before. In this case, and in cases like it, the self-regulatory processes are functioning normally, and so there is no self-regulation failure involved.

A second feature of a characterization of self-regulation failure with a progress-based criterion is that it accommodates not only "attainment ends" but also "maintenance ends" (Brodscholl, Kober, and Higgins, 2007). The difference is explained by Brodscholl et al. (2007) as follows.

> Within both attainment and maintenance, a desired, positive end-state is the reference point or comparison standard for the individual's goal pursuit. Moreover, both attainment and maintenance relate to positive outcomes (to be attained or maintained). What varies between these two goal pursuit conditions is simply the position of the individual's current state in relation to his or her desired end-state: The current state is discrepant with the desired end-state in attainment, but congruent to it in maintenance. (Brodscholl et al., 2007, p. 629)

Examples of maintenance ends are wanting to maintain one's current weight, or wanting to maintain one's current romantic relationship. Ends like these can be accommodated because the conception allows for the discrepancy to remain the same as long as that is a result of well-functioning self-regulatory processes. Moreover, when success is relative to the discrepancy, maintenance can naturally be considered self-regulation success when the discrepancy has already been minimized to be within an acceptable range (e.g., weighing 0.5 kg more or less than one's target weight). And should a breakdown of processes result in a discrepancy outside the acceptable range (e.g., gaining a couple of pounds, experiencing friction in the relationship), then that breakdown qualifies again as self-regulation failure.

A third and final feature is that breakdowns of processes that do not result in an increased discrepancy do not qualify as instances of self-regulation failure. On first sight, this may seem like a bug rather a feature, as it is somewhat counterintuitive to disconnect the notions of a breakdown or "non-normal functioning" or malfunction with all

their negative connotations from the notion of failure. However, it is important to see that problems with individual processes are not to be identified with a breakdown of the system that the conglomeration of processes together constitute. Saying that breakdowns of processes are sometimes not identifiable with breakdowns of the system is therefore not at all the same as flat-out denying that something went wrong. Any breakdown can point to weaknesses in individual aspects of the self-regulatory system. However, the reason why the disconnect is a feature and not a bug is that it allows for a certain *robustness* of the self-regulatory system as a whole, by having "backup mechanisms" play a role in one's self-regulation.

As an example, consider the person who wants to write a book, but who can never muster the self-control on his own to sit down and write when the opportunity arises. In those moments, there are clear, identifiable breakdowns in some of his self-motivating self-regulation processes. However, if this person ends up writing anyway because he has pledged to donate money to a cause he vehemently opposes if he does not write (see, for example, `http://www.stickk.com/`), he ends up decreasing the discrepancy and moving towards his end, even though breakdowns occurred. On the whole, then, his self-regulatory system did not fail with regard to the overall end.

Similarly, in a case of maintenance, consider the recovering alcoholic who has locked away his liquor in a cabinet and has given the key to his neighbor. If his immediate self-control breaks down, and he desperately and actively looks for a drink, there is a breakdown of his self-regulatory processes. However, if he ends up not drinking because he had increased the threshold for accessing his liquor, his overall self-regulatory system can be said to have worked (cf. Heath and Anderson (2010)).

In extension, the disconnect between malfunctioning processes and self-regulation failure even allows for breakdowns that result in a decrease of the discrepancy, such that the breakdown could be called functional in some way. While I suspect that these cases will be rare, we may suppose that the following contrived example could occur. A person wants to improve his overall physical fitness and, in order to reach that end, decides he will go on walks every evening in order to reach 10.000 steps, which he records on his fitness tracker. Then, one evening, he cannot muster the self-control to go out for his evening

walk and decides to skip it in favor of an evening of going out clubbing with his friends. At the end of the night, he still feels guilty about skipping his evening walk, until he checks his fitness tracker and finds, much to his surprise, that with all the dancing he had done he had recorded over 12.000 steps. In effect, then, like the previous evenings in which he went on his evening walk, this evening had brought him closer to his end of improving his physical fitness.

Notice that in the this scenario, the decrease of the discrepancy was the direct result of the breakdown, and not, as in previous examples, due to the workings of backup mechanisms. The question that arises now, however, is whether we should attribute this person with self-regulation failure for skipping his walk or self-regulation success for meeting his desired step count. The answer, according to my conception of success and failure, is neither. First, because the decrease in the discrepancy is not a result of normal functioning, attaining 12.000 steps cannot count as self-regulation success. Second, because the breakdown does not result in an increase of the discrepancy, this scenario does not qualify as failure either. Given that this person's behavior did get him closer to his end (so no failure), but without the behavior being guided in the right way by his end (so no success), I think this is an intuitively acceptable position.

As a final remark, notice that because the conception defines failure and success locally to the current iteration of the feedback loop, it is compatible with the idea that failures can have a positive effect to the overall pursuit of one's end. Sometimes, this effect may be quite direct, for example if a failure to perform a certain action ends up providing the person with some much needed "slack time" (for more on slack, see Mullainathan and Shafir (2013)), which results in higher productivity or better performance the following day. At other times, failures can have a positive overall effect on the temporally distal outcome through more indirect pathways of learning. This is especially the case with learning new skills (consider again learning to throw free throws in basketball where misses can be informative as to what needs to change or improve) but also in cases where the failures lead to new insights about oneself. For instance, failures to stick to one's diet can be very informative as to the circumstances under which the failures occur (e.g., at parties where food is in abundance and social norms are at play). Understanding one's own weaknesses and pitfalls is pivotal for

choosing the right self-regulation strategies (e.g., guarding oneself from temptation by making Holtonian resolutions; see Section 1.1) and failures can help identify what those weaknesses and pitfalls are. In this regard, having the opportunity to fail and being allowed to fail is important with regard to improving one's self-regulation.

Let us now take stock. I have given a characterization of self-regulation, as well as an account of self-regulation success and self-regulation failure. In addition, I have tried to show that the features of this account make it plausible that conceptualizing success and failure locally in relation to the current iteration of the feedback loop is a fruitful approach. Sticking with this conception, then, let us return to my characterization of self-regulation itself. As stated, the characterization I have given was in terms of its constitutive components. However, this characterization leaves out many aspects to self-regulation that may not be constitutive of the concept, but that are relevant for understanding self-regulation of human agents in the actual world and the various factors that self-regulatory success is dependent on. Such aspects include, but are not limited to, the following.

a Self-regulatory processes can be deliberate or automatic processes (e.g., Karoly, 1993; Bargh, Gollwitzer, Lee-Chai, Barndollar, and Trötschel, 2001; Bargh and Williams, 2006; Carver, Johnson, and Joormann, 2009; Milyavskaya and Inzlicht, 2017);

b Self-regulation is typically concerned with behavior, but not exclusively so (cf. mood regulation [e.g., Mayer and Gaschke (1988); Thayer, Newman, and McClain (1994); Schutz and Davis (2000)] or regulation of thoughts and "cognitions" [e.g.,Logan and Cowan (1984); Barutchu, Carter, Hester, and Levy (2013)]);

c Self-regulatory success often depends on one's ability to inhibit impulses (*self-control*; e.g., Baumeister and Heatherton (1996); Tice, Bratslavsky, and Baumeister (2001); see Fujita (2011) for an alternative conceptualization of self-control);

d Self-regulatory success often depends on one's confidence with respect to one's capabilities to meet situational demands (*perceived self-efficacy*; e.g., Wood and Bandura (1989); Bandura (1991, 2005));

e Self-regulatory success depends to some extent on personality traits and corresponding "self-regulatory style" such as being

impatient or laid back (e.g., Hoyle, 2006; Baumeister, Gailliot, DeWall, and Oaten, 2006);

f Self-regulatory success often depends on specific situational and environmental factors (e.g., one's social environment) and one's ability to foresee, adapt to, and/or modify relevant features of the environment or situation (Heath and Anderson, 2010; Levy, 2012; Torma, Aschemann-Witzel, and Thøgersen, 2018);

g Self-regulatory success often depends on one's abilities to form means-end coherent plans and sub-plans to attain the chosen end (Taylor, Pham, Rivkin, and Armor, 1998; Zimmerman, 2000; Mann, De Ridder, and Fujita, 2013; Fishbach and Hofmann, 2015).

In the remainder of this section, I will more fully discuss the core components and indicate with examples the different points at which self-regulation can break down. Where appropriate, I will mention how the contingent aspects that I have identified can play a role in both the maintenance and breakdown of self-regulation.

1.2.1 The Discrepancy-Reducing Feedback Loop

The notion of a feedback loop was introduced into the self-regulation literature by Carver and Scheier (Carver and Scheier, 1981, 1982). They explain the basis of a feedback loop as follows.

> [An] *input function* is the sensing of a present condition. That perception is then compared against a point of reference via a mechanism called a *comparator*. If a discrepancy is perceived between the present state and the reference value, a behavior is performed (*output function*), the goal of which is to reduce the discrepancy. The behavior does not counter the discrepancy directly but by having an impact on the system's environment (i.e., anything external to the system). Such an impact creates a change in the present condition, leading to a different perception, which in turn is compared anew with the reference value. This arrangement thus constitutes a closed loop of control, the overall purpose of which is to minimize deviations from the standard of comparison.

Carver and Scheier have suggested that such feedback loops "have a great deal to recommend them as a model of human functioning" (Carver and Scheier, 1982, p. 111). This idea was subsequently adopted by various authors and integrated into different models of self-regulation. In the context of education, for example, Zimmerman has theorized that "self-regulated learners" go through three iterative phases, viz. a *forethought phase* that involves processes of task analysis and self-motivation, a *performance phase* that involves processes of self-observation and self-control, and a *self-reflection phase* that involves self-judgment and self-reaction (Zimmerman, 2000, 2008). This comprehensive model captures a great deal of the complexity involved in self-regulated learning, but as a model of self-regulation itself, it is perhaps overly inclusive in taking on board aspects of agency such as self-reflection that are better understood as part of personal autonomy or self-governance. I will say more about this relationship between self-regulation and self-governance in Section 4.1.

My aim here is only to give an overview of self-regulation in order to provide a clearer picture of the various ways in which self-regulation can break down. For this, a more minimal conception that stays closer to Carver and Scheier's explanation of the feedback loop is more suitable. Such a conception can be found in the context of health, where the notion of a self-regulation feedback loop has been brought to the foreground through work by Baumeister and colleagues (Baumeister, Heatherton, and Tice, 1994; Baumeister and Heatherton, 1996; Muraven and Baumeister, 2000; Baumeister and Vohs, 2007; Vohs and Baumeister, 2016). On this view, three distinct components play a role in self-regulation: having standards, having a capacity to monitor oneself in relation to these standards, and having a capacity to direct oneself in order to change course if necessary. As these components appear to form the core of the self-regulatory feedback loop, I will discuss them in order and illustrate the different ways in which failures can occur in relation to each of them.

The standards one strives to meet

Having *standards* relates to having what Carver and Scheier called a "point of reference" to which perceptions can be compared. Importantly, as I specified in my characterization of self-regulation, these standards have to be the agent's own, in the minimal sense that the

agent considers it problematic, all things being equal, when the agent falls short of the standard. In agents who are not just self-regulating but also meet the more demanding standards of self-governance, many of the agents' standards will likely be self-endorsed in the strong, Frankfurtian sense of "wholeheartedly identifying" with the standard (Frankfurt, 1987), but having standards does not presuppose self-governance. All that is required on my view of self-regulation, is that the agent recognizes the standard as one's own, even if, when one were to stop and think about it, one would want to reconsider the standard (e.g., one might recognize one's urges for chocolate as one's own, even though, on reflection, one would wish that these urges would not drive one's behavior as much as it does). There are thorny issues to consider in relation to the subject of reflective endorsement and (non-)alienation—see Christman (2009, ch.7) for a discussion—but I will leave those to the side. For I only mean to distinguish between what from the agent's perspective are standards, and what from an outsider's perspective can be considered standards.

In the literature, these two perspectives are sometimes muddled together, but it is important to separate the two. One reason for this is that the difference has practical implications for interventions, as people's motivations will vary depending on whether a standard is one they consider (in some sense) their own or one that is externally imposed. More importantly, however, the difference highlights that self-regulation by itself is not connected to rules of morality or prudence or convention. For example, the agent who genuinely adopts the goal to drink oneself into a stupor and proceeds to do so successfully by monitoring one's level of inebriation and ordering more alcohol while still able to do so, can be said to be self-regulating, even if, from a third-person perspective, one might frown upon this behavior. Similarly, the person who genuinely does not care about following the rules of law (and even identifies with being an outlaw) can exhibit all kinds of behavior incongruent with societal standards without exhibiting self-regulation failure. In this regard, I deviate from some other perspectives on self-regulation, as reflected by this quote from developmental psychologist Kopp (1982), who writes that "[...] it is generally agreed that self-regulation demands awareness of socially approved behaviors and thus represents a significant aspect of the socialization of children" (Kopp, 1982, p. 200). Of course, the

development of self-regulation skills may go hand in hand with developing a moral compass, but to say that self-regulation demands such awareness is, on my view, too strong.

Let us now turn to the question of what standards are. In principle, many different constructs could play the role of standards in feedback loops in general, and this is also reflected in the characterization of standards given by Baumeister and Heatherton, who write that standards are "ideals, goals, or other conceptions of possible states" (Baumeister and Heatherton, 1996, p. 2). The inclusivity of such a characterization allows self-regulation theory to accommodate different approaches to operationalizing standards. For example, some researchers might take standards to mean judgments about what to do, while others interpret standards as plan-like states such as intentions.

In practice, the standards on which many self-regulation researchers focus are goals (e.g., Zimmerman and Kitsantas, 1997; Latham and Locke, 1991) and goal intentions (e.g., Sheeran, Webb, and Gollwitzer, 2005; Webb and Sheeran, 2007). For the purposes of this dissertation, I would like to follow suit and use these terms, as they intuitively correspond well to the notions of "ends" and "plans" respectively, as found in the planning theory of agency that I will be using in later chapters. Unfortunately, there is some contention in the literature about how these concepts of goals and goal intentions are related, which necessitates a brief interlude to clarify exactly how I will understand the terms.

In the psychological literature, authors typically think of goals as "internal representations of desired states" (Austin and Vancouver, 1996). This, I believe, is considered common ground. The contention, however, is about where to locate the agent's commitment towards a desired end. The way I interpret the debate, there are two camps. On the one hand, there are authors like Mann et al. (2013) who argue that goals are constituted by a goal intention plus a commitment to act:

> Although people may desire or intend to attain some outcome, they are not committed to that as a goal until they are willing to invest affect, cognition, and behavior in attaining it. Whereas goal intentions specify a desired end state, goal commitment indicates how much that end state is desired and motivates action. Merely having an intention is thus insufficient to constitute a goal. (Mann et al., 2013, p. 488)

On the other hand, there are those who think that goal intentions always involve a degree of commitment by virtue of being intentions rather than "mere" desires. Gollwitzer and Sheeran for example take goal intentions to be "self-instructions to attain certain outcomes or perform particular behaviors" that "imply a commitment to act that may vary in strength" (Gollwitzer and Sheeran, 2006, p. 70).

At first glance, the former position may seem confused in light of how the term "intention" is typically understood in philosophy, namely as a state one of whose distinctive characteristics is that it involves (some degree of) commitment to act. However, the position can be made sense of by taking into account that intention also has a colloquial use, such as when one utters a sentence like "on good intentions alone you cannot pay the mortgage". Here, the reference to "good intentions" is suggestive of a kind of wishful thinking, where there is not and never has been any real commitment to act. This understanding of the word "intention" however differs substantially from the theory-laden way it is understood in theories of the mind—both in psychology (e.g., The Theory of Planned Behavior (Ajzen, 1991)) and philosophy (e.g., The Theory of Planning Agency (Bratman, 1987; Holton, 2009))—that has informed the latter position.

To move forward from this debate, let me stipulate the way I will treat these concepts. I will understand goals as the set of propositions that describe end-states of agential doings such as the goal "to run a marathon" or the goal "to get married". In other words, they are propositions that the agent seeks to make true in the world. Typically, agential doings will be actions, like the examples just mentioned, but I use the slight contrived phrase "agential doings" to make room conceptually for mood-related or cognition-related goals as well, such as the goal "to be happy" or the goal "to make less hasty decisions," respectively.

Provided that goals are a specific set of propositions, I then take goal intentions to be a type of mental state that imply a commitment to act of varying strength in service of attaining a specified end. For example, having a goal intention "to run a marathon" can be seen as what Gollwitzer and Sheeran call a "self-instruction" to invest effort in preparing for and actually running a marathon. Put in Bratman's terms, goal intentions are a type of planning state.[5] In Chapter 4, I will

[5]Holton implicitly also takes this view; see Holton (2009, p. 8).

say more about how planning states should be understood, but here it suffices to say that states such as goal intentions involve a degree of commitment, are typically stable over time, and possess a degree of resistance against challenges (e.g., Sheeran and Abraham, 2003; Cooke and Sheeran, 2004). Moreover, goal intentions and other planning states are typically *partial*, in the sense that they lack the specificity to be actionable right away. Instead, they provide a structure within which the agent can subsequently—right away or at a later time—form so-called "behavioral intentions" that specify concrete actions to take in service of fulfilling one's goal intention.

With the conceptions of goals and goal intentions now fixed, I can also say what it means on my account to "set" or "adopt" a goal (a topic with an extensive literature on its own, see for example Locke, Shaw, Saari, and Latham (1981); Locke and Latham (2002); Hurn, Kneebone, and Cropley (2006)). On my view, adopting a goal just is forming a goal intention (also known as a Bratmanian plan). Subsequently, having a goal intention brings into play norms of means-end coherence and consistency that I will also discuss more fully in Chapter 4.

This concludes the interlude. Now that we have a grasp on what it means to have standards, let us look at the different ways in which problems with standards can lead to self-regulation failure. First, some authors, like Hofmann, Schmeichel, and Baddeley suggest that "[p]eople may fail at self-regulation owing to a lack of standards" (Hofmann et al., 2012, p. 174). A similar thought can be found in Baumeister and Heatherton (1996), where the authors write that "a lack of standards [...] can prevent effective self-regulation" (Baumeister and Heatherton, 1996, p. 2).

Without any qualification, this line of thinking strikes me as odd. For if there are no standards at all to compare perceptions against, there is no closed feedback loop and therefore no self-regulation to speak of. What these authors are likely getting at is that problems may occur when there are no *clear* standards. For example, when one decides to go to bed "on time" without specifying either a specific time or an observable event for when to go to bed (e.g., the sun going down), then there is an obvious risk of going to bed at a time that one will regret the next morning due to a resulting sleep deficit.[6]

[6]Notice that the point here is not that everyone should always adhere to strict bedtimes. Rather, the example shows that one can fail, by one's own lights, to self-regulate

Another interpretation of what authors might mean by "a lack of standards" is that an overall standard for the end does not specify standards for the means. For example, having a goal intention "to lose body fat" does not specify by itself which foods to consume or what exercises to perform in order to lose body fat. As such, individuals who have formed such a goal intention but have not worked out how to achieve the end, may fail at their self-regulation by consuming too many calories simply because they do not know how many calories they should eat per day, or how many calories certain foods contain.

What the examples point to is that self-regulation often requires of agents that they form goal intentions that are "sufficiently clear". Clarity in this sense can be conceived of along different dimensions such as specificity, measurability, achievability, realism and timeliness (*SMART* goals; Doran (1981); MacLeod (2012)), but sufficiency often depends on the individual and the circumstances. I will not attempt any generalization of it here. The point, rather, is that forming insufficiently clear goal intentions can be a way to hamper one's self-regulation.

Second, problems with standards may also occur when there are conflicting, incompatible standards. For example, one's intention to go to bed at 11 p.m. might conflict with one's intention to respond to all student inquires about an upcoming exam. Similarly, but on a different level, one's goal intention to stick to a grueling diet and exercise regime might at times clash with one's values about living a rich and full life.

Third, self-regulation problems can arise when the standards do not correspond to the perceptions one is processing. For this point, however, I will first turn to explaining self-monitoring.

Self-monitoring

Against the background of standards, self-regulation involves *self-monitoring*: keeping track of what one is doing and how it aligns with one's standards. Baumeister and Heatherton (1996) call this the "test phase" of self-regulation, as it involves "comparing the actual state of the self to the standards" (Baumeister and Heatherton, 1996, p. 2). If the comparison is favorable—in the sense that the discrepancy between

by not having clear enough standards to compare one's perceptions against.

the current observation and the standard has reduced relative to the previous perception—then no intervention from the self is required. Otherwise, self-direction is needed to course-correct oneself (see the next subsection).

Comparing the actual state to a standard implies having information about the actual state, and as such, self-monitoring really has two separate aspects, viz. *self-observation* and *self-assessment*. The former notion should be understood broadly to include not just being aware of one's bodily movements, but also noticing feelings of discomfort, gnawing suspicions, etc. Moreover, self-observation occurs in relation to context and time, which is why losing track of time can be problematic for self-regulation if one's end is time-sensitive.

With regard to the relation between self-observation and self-assessment, it is worth noting the importance of making sure that there is a match between the kinds of observational inputs one seeks and the kinds of observational inputs that are needed for the purposes of self-assessment in relation to the relevant standard. For standards can specify either specific behaviors or specific outcomes of behaviors, and this affects the kinds of observational inputs that are relevant (Harkin, Webb, Chang, Prestwich, Conner, Kellar, and Sheeran, 2015). For example, a standard such as a goal intention "to win more tennis matches" implies that relevant observations will concern the number of matches played and the number of matches won. This is not to say that observing the manner in which one played is not relevant for future matches, because it is. However, it does mean that only observing for example how technically sound one's technique was during the match is not the right kind of observation. However, if the goal intention had been "to become a technically better tennis player," then the kind of observations that would be relevant would have been reversed: observing one technique would be much more relevant than the outcome of the match.

Moreover, even if one were to focus on either a specific behavior or a specific behavioral outcome, it is still important that there is a match between the kind of observation one makes and the relevant standard. Consider for example learning to hit a topspin forehand in tennis, and suppose that one focuses solely on the behavior itself. In this scenario, there can still be a mismatch if one observes only the swing path of the racket without perceiving, through proprioception, the (lack of)

turning of the shoulders and driving of the legs. Similarly, were one to focus solely on the outcome of the behavior, one could miss crucial feedback by only looking at the trajectory of the ball without listening to the sound of the strings on impact.

Mismatches like these can contribute to self-regulation failure when one is monitoring only a subset of the relevant aspects to a behavior or outcome of a behavior. For another example of this, consider someone who wants to lose 5% body fat, but who monitors one's body weight instead of fat percentage and diligently tracks one's progression of an exercise regime without monitoring one's food intake. While this person might succeed, chances are that he will not.

A different way in which issues with monitoring can lead to self-regulation failure is when one only checks one's progress when one already knows it is going to be positive. For example, some people will only measure their weight while on a diet or exercise regime, but not when they suspect that there is no progress to see. Similarly, one may only look at one's pedometer if one already has a sense that it is going to confirm that their goal for the day has been met. In some instances, this "positive bias monitoring" might be indicative of a deliberate strategy to protect one's own self-image by looking away when things are bad, and patting oneself on the back when there is progress to be seen. In other instances, though, this type of monitoring may be a genuine form of positive self-confirmation, where one already has a general sense of moving in the right direction, and one is looking for an extra boost of motivation. However, even in the instances of the latter kind, there is a potential pitfall with this way of monitoring with regard to self-regulation in that people who do this are prone to miss relevant information precisely at moments when they are moving away from their intended end.

Yet another way in which monitoring can negatively affect self-regulation is by monitoring a semi-automatic process too closely. For example, for most people, writing one's signature is a fluent and near-automatic movement. However, when one focuses closely on the task, for instance when one is put under pressure to produce a flawless signature on a formal document, it can become difficult to keep the motion going smoothly. Similarly, when one pays close attention to one's breathing, it can begin to feel forced.

Thus far, I have focused on monitoring aspects of the self for the

direct purpose of comparison with the standard. As a final remark about monitoring and its role in self-regulation, consider that self-monitoring also involves keeping track of one's cognitive, motivational and volitional resources. This aspect of self-monitoring is less about the standards, and more about self-direction and assessing how much effort one can expend on course-correcting oneself given the circumstances. Such assessments are essential for choosing appropriate self-interventions, such as making plans or calling a friend for support. Importantly, these assessments are not only relevant for present situations, but can also play a role in deliberating about future situations. For example, if one observes a decline in motivational resources for writing a manuscript, one can schedule an appointment with one's coauthors for either a collaborate writing session or a thorough review of one's work so far (the metaphorical carrot and the stick, respectively). This role that self-monitoring plays in the temporal dimension of self-regulation is sometimes overlooked, but it is good to keep it in mind in the next section about self-direction.

Self-direction

Finally, the third core component of self-regulation is having a means of *directing* oneself in order to actively course-correct if necessary. Typically, this means altering one's behavior to better meet the standards, but it can also mean altering one's thoughts or mood in relation to a standard. While it may perhaps sound a little contrived to speak of self-directed mood, there is evidence that mood is controllable to a certain extent by engaging in particular activities such as exercise (e.g., Thayer et al., 1994), and I will be assuming that mood can indeed be directed in some (potentially indirect) way. In what follows, though, I will focus on behavior rather than thought or mood for reasons of convenience, but in principle, it should be possible to substitute behavior with either thought or mood.

Another point I want to draw attention to upfront is that I will not discuss in this subsection the kind of self-assessment one does when one reflects on a higher level on whether one even wants to engage in self-direction, such as when one reconsiders whether one is still committed to striving towards a particular end. As I will make clear in later chapters, such reflections are key to human agency, but I tend to think of those kinds of reflections as belonging to processes of

self-governance, rather than self-regulation. Similarly, I am inclined to think that high-level goal setting (i.e. plan formation) is best understood as part of self-governance and not self-regulation (but see Mann et al. (2013) for a dissenting opinion). Self-direction as I will be discussing it here should be understood as local to the self-regulatory processes that have already been initiated. Clearly, there is an interplay between processes of self-regulation and processes of self-governance, and I will say more about this in Section 4.1, but I think these concepts should be considered distinct from one another.

What is constitutive of self-directed behavior is that the behavior is in a meaningful sense "linked up" with the standards, so that when the agent's behavior aligns with the agent's standards, we can attribute self-regulation success to the agent. In contrast, the agent who finds her behavior aligned with her standards by mere happenstance is not successful in her self-regulation, just lucky. To illustrate this point, consider the following two parallel cases of an agent who is both lonely and unhappy with her own weight, and who, after deliberation, adopts goal intentions to take steps to deal with each of these issues.

> *Self-directed weight loss.* The agent decides to get a dog to increase her exercise in order to get into better shape. After a month, she has lost weight by her going on frequent walks with her dog.
>
> *Lucky weight loss.* The agent decides to get a dog for companionship. After a month, she is surprised to discover that she lost weight by her going on frequent walks with her dog.

In the first variant of the scenario, self-regulation success can be attributed to the agent since her decision to get a dog was a result of her deliberation about how to lose weight, and her subsequent dog-walking was in service of that same goal intention. Notice how this is not the same as saying that the agent must be aware of her goal intention to lose weight each time she walks the dog. As for example Milyavskaya and Inzlicht (2017) have recently emphasized again, self-regulation also includes "effortless, automatic, or habitual forms of goal-directed behavior" (Milyavskaya and Inzlicht, 2017, p. 11), and as such, having an established dog-walking routine can play an integral part in the agent's self-regulation success. Rather, the point is that *de facto* alignment, while necessary, is not sufficient for self-regulation: there must also be a particular kind of connection between the goal

intention and the subsequent behaviors. The contrast of the second variant of the scenario makes this clear. In this case, the agent losing weight is mere happenstance. Though losing weight is aligned with her standards (i.e. her goal intention to lose weight), her losing weight is not, strictly speaking, anything she can be credited for. Though of course she was involved in the process, the behaviors that caused her to lose weight were not connected in the right way to her goal intention to lose weight.[7]

Let us now turn our attention to how self-direction is typically understood. In the literature on self-regulation, self-direction is predominantly operationalized in terms of exerting *self-control*, which is often considered as the effortful inhibition of impulses (Baumeister and Heatherton, 1996; Baumeister and Vohs, 2007). On this view, steering oneself back in the direction of one's standards involves overriding impulses or urges of varying strengths that are pulling in a direction that is incongruent with one's standards (cf. dual-system theories in social psychology (e.g., Chaiken and Trope, 1999; Frankish, 2010)). The picture is essentially one of *self-restraint*: one has to resolve an internal conflict about which course of action to take by using cognitive and motivational resources to quell the impulse and thereby firmly standing one's ground in light of temptation. Examples of cases involving self-restraint like this abound: declining a tempting chocolate cake at a party, not taking a much-craved cigarette when offered, saying no to unprotected sex in the heat of the moment, etc.

It is worth noting, however, that even for these cases that are supposed to exemplify impulse inhibition, it actually is not evident without any further information about the cases whether they indeed involve effortful impulse inhibition. For example, if the act of declining a chocolate cake at a party was a semi-automatic response resulting from a predetermined action plan about what to say when cake would be offered, then the internal conflict would have been resolved via a different mechanism than effortfully inhibiting the impulse to reach for the cake in the moment. This point has been emphasized by Kentaro Fujita (2011) who has argued that the term self-control is better understood more broadly as "the resolution of a dual-motive conflict"

[7] Cases such as these mirror so-called "causal deviant chain" cases in which agents have desires and beliefs that make it reasonable to ϕ, that cause one to ϕ, but where the ϕ-ing is not intentional (Davidson, 1973; Peacocke, 1979).

between proximal and distal motivations (Fujita, 2011, p. 352) and that the effortful inhibition of impulses is only one possible mechanism available to agents to achieve such a resolution. This is a noteworthy move, because it creates conceptual space in the discourse about self-control to discuss alternative mechanisms that people utilize or may utilize in resolving their dual-motive conflicts. Such mechanisms may include, but are not limited to, regulating the availability and opportunity to indulge in temptation, utilizing planning strategies to guard themselves from temptations (cf. Holtonian resolutions) or engaging in cognitive reconstrual where agents reconstrue temptations in more abstract terms in order to look beyond the immediate reward they offer (e.g., Fujita, 2008). Having such mechanisms in view is essential for designing support structures such as self-regulation support systems that are able to offer appropriate interventions and insights.

However, even with this broader conceptualization of self-control in place, one must realize that self-control is not synonymous with self-direction, as self-direction also includes forms in which there is no internal conflict between proximal and distal motivations. I will mention two of such forms here, but I suspect others forms could be identified as well. The first is a form I like to call *self-correction* that involves correcting certain innate or habitual tendencies that one might have in a "cool," deliberate way (as opposed to "hot" impulse control). With regard to innate tendencies, we can turn to the extensive literature about the various ways in which people are biased and "predictably irrational" (Ariely, 2008). A discussion of the various pathways that lead to to predictably irrational behavior is beyond the scope of this chapter, but the point I want to get across is that there are many cases in which people do experience a pull into a direction incongruent with one's goal, but where there are no resulting internal conflict between two motivations. Rather, the pull one experiences simply poses a challenge or obstacle to overcome.

As an example, consider a person who has run out of laundry detergent and goes to the store to buy more. She always uses the store-brand detergent, but now that she is in the store, she is faced with a big and flashy commercial for a "buy 2, get 1 free" deal on a more expensive A-brand detergent. Suppose that the person in this scenario is enticed to the point where she entertains the idea of buying the A-brand detergent. She checks the price of the A-brand detergent and calculates

that with the "2+1 deal" each bottle of A-brand detergent is now only slightly more expensive than her regular, store-brand detergent, but not by much.[8] Looking at both products, she deliberates about what to do. There is no internal conflict between opposing motivations, because whichever product she picks will satisfy her need for laundry detergent, but she does have a choice to make. However, in her deliberations, it occurs to her that the commercial is playing into her bias, and so, despite feeling the pull of her bias towards free stuff, she self-corrects and picks up her regular, store-brand detergent.

Similarly, habitual tendencies can also pose obstacles without representing an opposing motivation. For example, consider a person who, in an attempt to aid his goal striving towards weight loss, decides to take a longer route home from work than his usual route. This new route requires going right at the first intersection instead of left. Because this person is so used to taking his old route, he frequently finds himself automatically switching lanes while lost in thought about work, readying himself to turn left, only to realize that he needs to go right. In this case, he will have to self-correct in order to overcome his habitual tendency to go left.

A confounding issue with separating different forms of self-direction like this is that problems with self-regulation will frequently require multiple forms of self-direction to work in tandem. For example, going through the process of self-correction might lay bare an emotionally charged conflict between opposing motivations that requires self-restraint to resolve. As an example, consider a case of hyperbolic discounting (Ainslie, 2001) where someone routinely puts off making a decision about buying instead of renting a house. Suppose this person hyperbolically discounts the financial gains he stands to make if he were to buy a house, and instead opts to keep renting in light of the immediate payoff of living somewhere familiar and not having to go through any hassle (not having to move, but also not having paperwork to deal with since the rental agreement is already in place). Now, if this person were to learn about hyperbolic discounting, he might recognize this pattern in his own thinking, and decide to self-correct by giving more weight to the future financial gains he stands to make if he buys. Initially, this is a straightforward instance of self-correction. However,

[8] For simplicity, let us assume that both products are the same in volume, or at least can be used for the same number of laundry batches.

suppose that once he self-corrects, and he realizes that buying a house is in his best interest in the long term, he subsequently finds himself confronted with fears about the responsibilities that come with having a mortgage (e.g., fears about job security). If he is still convinced that he wants to go through with buying a house because he believes it to be in his best long-term interest, he will now have to appeal to a different mechanism (e.g., self-restraint) to overcome the inclination to keep renting, which on this deeper layer of the scenario can now be considered a proximal temptation.[9]

Notwithstanding such intricacies of the interplay between forms of self-direction, dissecting self-direction into separate forms has the advantage of getting a clearer picture of how different forms are likely to require different cognitive and affective capacities. While I will not presume to know exactly how each form relates to specific neuropsychological pathways (but see Braver (2012); Braver, Krug, Chiew, Kool, Westbrook, Clement, Adcock, Barch, Botvinick, Carver, Cools, Custers, Dickinson, Dweck, Fishbach, Gollwitzer, Hess, Isaacowitz, Mather, Murayama, Pessoa, Samanez-Larkin, and Somerville (2014)), my point here is just that more nuanced thinking about these forms can lay bare subtle differences that may be relevant for offering effective support.

The second form of self-direction that I want to mention besides self-control is that of *self-guidance*, where there really is no internal conflict to resolve, and the steering occurs in response to an observed discrepancy that occurred as a result of bad luck, incomplete or incorrect knowledge, or because of malfunctioning processes of self-monitoring. As an example of bad luck, consider someone trying to leave the building where he works from a side exit, which would put him closer to the public transportation that he needs to take to get home. Supposing his end is indeed to get home, this is a reasonable means to achieve his end. However, if the side exit happens to be closed (for reasons unbeknownst to him), the discrepancy between him and his end has been increased instead of decreased. Recall that

[9] Notice how differentiating between self-restraint and self-correction is different from categorizing impulses, tendencies and urges in terms of how difficult they are to suppress. Such an approach, that one might try to spell out in terms of weaker impulses needing "merely" self-correction and stronger impulses needing full-blown self-restraint, would face various problems, most obviously with explaining how some instances of self-correction are extremely difficult to overcome because the bias involved is particularly persistent.

this does not count as self-regulation failure though, as that would require a malfunctioning of processes, which is not the case. However, he does have to determine a different means to his end and direct himself to leave the building via the main exit, which he does. In this instance, there is no conflict of motivations, there are no impulses to inhibit, but there is self-direction nonetheless.

As an example of incomplete or incorrect knowledge, consider someone who is being made aware that being exposed to blue light late at night is disruptive to her ability to fall asleep. If we suppose that this is new information, and that having that particular kind of light switched on late at night is not intrinsically attractive to her, she may easily change course by switching to a different kind of light bulb.

As an illustration of malfunctioning self-monitoring, consider a cyclist who wants his heart rate to be within a certain range while he cycles. Suppose that during his training session his mind wanders, causing him to slow down just a bit, resulting in his heart rate dropping below his aim. Once the cyclist looks down at his watch to check his heart rate and he sees that it is too low in relation to his standard, he readily picks up the pace in order to bring his heart rate back up. In this situation, there was no proximal motivation that led him astray; it was just that he had not paid well-enough attention to his heart rate.

The observation that self-direction is more than just self-control is important for two reasons. First, having insight into what form of self-direction a situation requires stands to be relevant for designing self-regulation support systems that can effectively and appropriately help with getting back on track towards meeting one's ends. Such insights can serve as input into automated decision-making processes about what interventions are suitable at the present moment and in the given context (e.g., a subtle nudge versus an invitation to make a specific plan), but they perhaps may also be shared directly with the individual as advice about which strategy to adopt.

Second, beyond these practical advantages of having a more nuanced view of self-direction, the observation is also essential for breaking with a flawed and outdated view of agency in which human agents are portrayed as beings who are solely occupied with containing their urges, constraining their thoughts, and suppressing their appetites. Seeing that self-direction is more than self-control and self-restraint puts us in a position to also recognize and appreciate the ways in which

agency is partly constituted by our ability to determine for ourselves the paths in life we want to take. This aspect of agency deserves recognition in the discourse about self-regulation support, because the interventions that will be proposed will have to appropriately balanced with regard to this freedom we have to determine for ourselves how we want to live our lives. This, as I have mentioned before, is a subject we will return to in the upcoming chapters.

For now, let us consider how issues with self-direction can lead to self-regulation failure. Clearly, the ability to steer oneself in a certain direction lies at the heart of self-regulation, but because of its complexity, self-direction is also susceptible to breaking down through a variety of pathways. For example, numerous studies on impulse inhibition have shown that whether one is capable of doing so is dependent on how much effort one has already exerted. For example, Vohs, Baumeister, Schmeichel, Twenge, Nelson, and Tice (2008) showed that participants who had to make a series of choices (e.g., on consumer products or college course options) performed worse on subsequent tasks than did a control group whose participants where only asked to consider the products or options without making decisions about them (Vohs et al., 2008). The same authors also reported similar findings from a field study in which the self-reported degree of previous active decision making predicted how well they performed on subsequent self-control tasks. Another set of studies from Schmeichel, Vohs, and Baumeister (2003) found that participants who had already been taxed by previous decision making scored worse than a control group on measures of intellectual performance such as logical reasoning or reading comprehension tasks (Schmeichel et al., 2003).

In light of these and several similar findings, Baumeister and colleagues have suggested that the "strength" needed for self-restraint is a limited resource that can be depleted, and have likened this resource to a muscle that can be fatigued with use (e.g., Muraven and Baumeister, 2000). This conceptualization and corresponding analogy have since then been scrutinized by authors such as Inzlicht and Schmeichel (2012) who have argued that the available evidence does not point towards a resource model, but rather towards a process model of depletion where exerting self-restraint at time t_1 causes temporary shifts in motivation and attention that undermine self-restraint at a later time (Inzlicht and Schmeichel, 2012).

That motivation is at least part of the explanation was suggested by Muraven and Slessareva (2003), who found that if people had enough of an incentive, for example a monetary incentive or an altruistic incentive, they could overcome depletion effects (Muraven and Slessareva, 2003). Based on their findings, they suggested that motivation should be considered a moderator of effortful impulse inhibition, such that high motivation can help individuals, at least up to a certain point, to compensate for a depletion effect. For example, one may be too depleted to drag oneself from the couch and into bed, but find some energy in reserve when it turns out that a family member is in medical need.

While Inzlicht and Schmeichel have taken findings such as the ones from Muraven and Slessareva as evidence that depletion really is just a temporary shift in motivation that can be shifted back through incentives, Baumeister and Vohs have taken the same results on board by suggesting that many of the depletion effects that can be observed in the above-mentioned studies could also be a conservation mechanism aimed to conserve a partly depleted resource from depleting even further, but which can be overridden if motivation is high enough (Baumeister and Vohs, 2007). Despite this lack of consensus on how the phenomenon is best explained, however, there is consensus that the results of the studies suggest that effortfully inhibiting impulses affects people's ability or willingness to exhibit self-restraint at a later time (cf. also Kamphorst et al., 2018).

In addition, there is also evidence that personality traits influence one's ability to inhibit one's impulses (*trait self-control* or *dispositional self-control*; e.g., Friese and Hofmann (2009); De Ridder, Lensvelt-Mulders, Finkenauer, Stok, and Baumeister (2012); De Ridder and Gillebaart (2017)). For example, people who score high on the Self-Control Scale (Tangney, Baumeister, and Boone, 2004) are generally better equipped to inhibit their impulses than those with a lower score (e.g., Schmeichel and Zell, 2007). Moreover, personality traits also affect self-regulation success by influencing one's "self-regulation style": someone who is less likely to plan ahead might be more likely to be susceptible to temptation, simply because they did not engage in any kind of foresight. For instance, someone who is averse to making specific plans such as Holtonian resolutions to stay faithful to one's monogamous partner might find oneself more easily tempted to stray.

Besides one's willingness to plan, one's abilities to plan can also affect self-direction. Importantly, as authors such as Gollwitzer and Sheeran have emphasized, having formed plans in the past helps shape one's semi-automatic behaviors in the present because the behavior has to a certain extent already been specified (e.g., Sheeran et al., 2005; Gollwitzer and Sheeran, 2006). I will return to this subject in Section 1.2.2 when discussing interventions, but here I want to focus on the idea that there are individual differences in people's abilities (innate or learnt) to "self-program" oneself (cf. Slors (2015)). For example, some people might have developed a keen sense of their own weaknesses in particular situations, and can very acutely predict what threats particular situations will pose. As such, these people are in a better position to guard themselves from these threats by taking precautionary measures such as making specific if-then plans about how to respond to the threats. On the other hand, people who have a less developed capacity for making these predictions might find themselves exposed more often to unexpected temptations.

It is important not to understate the relevance of how well one's planning abilities are developed, as planning is an essential part of self-regulation. Mann et al. for example note that "goal striving refers to the process of *planning* and performing [...] behaviors necessary to achieve [...] goals" (my italics) (Mann et al., 2013, p. 491). Bandura puts it even more boldly:

> Most human behavior, being purposive, is regulated by forethought. The future time perspective manifests itself in many different ways. People form beliefs about what they can do, they anticipate the likely consequences of prospective actions, they set goals for themselves, and they otherwise plan courses of action that are likely to produce desired outcomes. Through exercise of forethought, people motivate themselves and guide their actions in an anticipatory proactive way. (Bandura, 1991, p. 248)

The emphasis on planning and the temporality of agency suggests that learning to make more effective plans or being supported in making effective plans might both be effective strategies to improve one's self-regulation. As we will later see, this idea has given rise to the development of various intervention techniques to do just that. However, as I have touched on before, it is essential that these techniques

be carefully scrutinized, since planning not only plays a big role in self-regulation (e.g., making plans about means to an end), but also plays an integral role in our agency and determining for ourselves our path in life (see Chapter 4).

I will introduce the intervention techniques in Section 1.2.2 and in Chapter 2. First, however, I will conclude this section by discussing two more ways in which self-direction can be supported or hampered. First, self-direction can also be helped or hindered via perceived self-efficacy. A good example of this can be found in the literature on procrastination. For a long time, procrastination was thought to be driven, at least for some people, by a fear of failure. However, initial studies focusing on fear-based procrastination failed to find a statistically significant relationship. Recently, however, Haghbin, McCaffrey, and Pychyl (2012) found that the relation between fear of failure and procrastination is actually moderated by perceived competence (a construct highly related to self-efficacy (Miserandino, 1996)), so that only those people who have a fear of failure *and* judge themselves as not competent in relation to a specific task are prone to procrastinate on that task (Haghbin et al., 2012).

Finally, self-direction is also affected by the context that a person finds oneself in. In the literature, the situation one finds oneself in is often described as a potential minefield of temptations that one can succumb to. We have already seen a number of examples of this. What for a long time has been overlooked, however, is the positive role that the environment can play in supporting one's self-regulatory efforts (Heath and Anderson, 2010; Levy, 2012). In recent years, the notion that the environment can be shaped and utilized for behavior change has gained a lot of traction among regulators and policymakers in discussions about potential nudges to change behavior (e.g., footsteps leading away from elevators to a staircase, healthier food items being placed more centrally in stores, see Thaler and Sunstein (2008); John, Smith, and Stoker (2009)), but *ecological engineering* (Levy, 2012) as a viable strategy for individuals to bolster their own self-regulation, remains less visible. This is unfortunate, as reshaping our environment can be a very effective way of lowering thresholds for certain behaviors. A good example comes from Heath and Anderson, who mention putting one's running gear next to one's bed. They argue that this lowers the threshold for going for a run, as it serves as a visual

reminder to go running, and has the practical advantage of no longer having to think about finding all the necessary items in the morning. Moreover, a clever use of the environment can also increase the threshold of engaging in a behavior, such as when one unplugs all the cords from one's computer in the evening, so that it becomes more difficult to give into the temptation of sitting behind the computer screen all night.

There remains much that can be said about each of these aspects of self-direction, but I will draw the section to a close here. With the foregoing discussion of self-direction, we have now discussed all three components and have closed the self-regulation feedback loop. For once one has steered oneself in the direction of one's end, a new iteration of the loop begins with more self-observation and self-assessment. As such, the basic picture of self-regulation is complete. The take-away of this section is that self-regulation failure is a broader concept than the classic and revisionist accounts of weakness of will (Section 1.1), but one that is more constrained than how it is sometimes portrayed in the literature. My hope is that this overview will serve as a scaffolding of sorts to help with understanding the ways in which self-regulation can be supported by the technologies that I will introduce in Chapter 2. That said, I also believe the arguments in the later chapters of this dissertation to be sufficiently robust to withstand some disagreement about the specifics of the picture of self-regulation that I have sketched in this section. Readers who have such disagreements are therefore encouraged to keep reading.

In the following subsection, I will discuss potential intervention strategies. Then, in Section 1.3, I will conclude the chapter with a discussion about the reasons people report both for developing technological interventions and for engaging with such technologies.

1.2.2 Interventions

In the past, various self-regulation interventions have been developed and tested. Some focus on increasing goal progress monitoring (e.g., Harkin et al., 2015), while others use mental visualization of both the positive aspects associated with a specific goal and the obstacles that stand in the way of attaining this goal (*mental contrasting*; e.g., Oettingen, Mayer, and Thorpe (2010)). One of the more well-established

strategies is to have people form highly specific if-then plans called *implementation intentions* (e.g., Gollwitzer, 1996). By coupling a salient cue from the environment with a predetermined action plan, people can "program themselves" to initiate desired action sequences semi-automatically without any redeliberation (Gollwitzer, 1993; Baumeister et al., 1994).

The strategy of making implementation intentions has proven its worth in a wide variety of settings, for example in increasing the frequency of self-examination for breast cancer screening (Orbell, Hodgkins, and Sheeran, 1997), increasing attendance for cervical cancer screening (Sheeran and Orbell, 2000), stimulating people to take their vitamin C tablets (Sheeran and Orbell, 1999), promoting shopping in a bio-shop (Bamberg, 2002), improving dietary quality (Adriaanse, Oettingen, Gollwitzer, Hennes, De Ridder, and De Wit, 2010), increasing plastic recycling (Holland, Aarts, and Langendam, 2006) and increasing job seeking activities (Hooft, Born, Taris, van der Flier, and Blonk, 2005). Implementation intentions are sometimes also used in concert with other methods such as mental contrasting (e.g., Adriaanse et al., 2010).

Saliently, Masicampo and Baumeister even showed that merely forming plans can help to eliminate the cognitive effects of unfulfilled goals such as having nagging, intrusive thoughts about activities or tasks that still need to be done (Masicampo and Baumeister, 2011). These empirical findings are in line with the theoretical contention by Bratman and others that making plans can "settle the matter" for a future time. Once the plan is made, one can direct more of one's cognitive resources to other pursuits, knowing that when the time comes, one knows what to do.

That planning interventions have had a measure of success is not surprising given the prevalent role that planning plays in self-regulation in general (see again Section 1.2.1). However, researchers are also finding that there is no such thing as a one-size-fits-all planning intervention. As we have seen, whether someone is able to make effective plans is not only influenced by one's personality traits but is also highly dependent on the specific goals people have and the contexts that people find themselves in. Likewise, the success of interventions that focus on monitoring are highly dependent on what one is monitoring and in which context. This makes designing robust interventions

for self-regulation support difficult, as the interventions need to be suited to the individual and be responsive to changing circumstances as well.

It is here that technology comes in. Recent advancements in technology are not only changing the ways in which existing intervention strategies are deployed (e.g., via the Internet) and goal progress is monitored (e.g., via sensor systems), but also the way in which the interventions themselves are being made to fit an individual's specific preferences and behaviors. The ubiquitous presence of smartphones in particular has had an impact on how interventions are performed, as they allow for continuous data collection of behavioral patterns and context via an array of built-in sensors (e.g., gyroscope, GPS, light, bluetooth low energy) and provide programmable alarms and notifications (that for example can be used as cues for implementation intentions). Moreover, they offer a platform for automated question-response interactions (commonly known as *ecological momentary assessment*), sending educational or motivational messages, and real-time chat with caretakers, clinicians, peers, or researchers. This infrastructure around smartphones is now widely used to deliver *persuasive health interventions*, both by researchers (e.g., Van de Ven, Henriques, Hoogendoorn, Klein, McGovern, Nelson, Silva, and Tousset, 2012; Klein, Mogles, and Van Wissen, 2014; Boh, Lemmens, Jansen, Nederkoorn, Kerkhofs, Spanakis, Weiss, and Roefs, 2016) and by commercial parties (e.g., "Goalie" by Sense Health).

Technology-based interventions are typically available around the clock, which allows for timely and, in combination with certain sensor data, context-appropriate interventions. At the moment, this *context awareness* of technological interventions is often limited to location-tracking mechanisms such as GPS. Going forward, however, this awareness is bound to be extended in various ways using sensors in clothes (*wearables*; for example to track biophysical measures such as perspiration), sensors in the ambient environment (e.g., tracking one's activities around the house; cf. Rashidi, Cook, Holder, and Schmitter-Edgecombe (2011); Frey (2013)), or proximity sensors in one's smartphone or smartwatch to detect social interactions (cf. Stolk (2015)).

More importantly, we are at the brink of a new stage of development of self-regulation interventions in which techniques from the field of Artificial Intelligence such as *machine learning algorithms* are

being brought into the domain of behavior change. These techniques allow for automatic drawing of inferences and hypotheses, making suggestions based on these inferences and hypotheses that the user has neither provided nor explicitly approved, continuously updating its domain knowledge, gathering information about a user's physical and social environment, learning about and being adaptive to a user's changing preferences, behavioral patterns and context, and providing therapeutic dialogues without human intervention (cf. Beun, Brinkman, Fitrianie, Griffioen-Both, Horsch, Lancee, and Spruit (2016)).

These techniques stand to further change how patients, clinicians and researchers will approach health and behavior change. Moreover, they will bring about a transformation of the landscape of self-regulation interventions, with transparent, tool-like behavior change systems making place for full-blown *e-coaching systems* that exhibit a relatively high level of sophistication and independence.

This heightened level of sophistication and independence will change how we relate to these systems and this brings with it a set of concerns that should be considered by developers and policymakers alike. In Chapter 3 I will elaborate on these concerns. First, however, it is good to understand the reasons that are being given for developing these technologies, as well as the reasons individuals may have for engaging with such technologies. It is this subject that will be discussed in the final section of this chapter.

1.3 Pro Tanto Reasons for Developing Technological Interventions Aimed at Reducing Self-Regulation Failure

In this section I will articulate different perspectives from which people draw motivation for wanting to develop technological interventions aimed at reducing self-regulation failure (or at least a subset of these failures). More specifically, I will discuss the two most prominent perspectives found in the literature, as well as an important third perspective that is sometimes overlooked outside of philosophy. While all three perspectives are in principle relevant to the more general matter of developing any type of intervention to reduce self-regulation failure, I will focus the discussion here on developing technological

interventions only. The aim of this section is twofold, namely to showcase the force of the arguments that are being brought forward in favor of these technologies, and, at the same time, to intervene in the discourse about these technologies by articulating explicitly that the reasons from all these perspectives are only *pro tanto* reasons that can be outweighed by more weighty reasons.

The first prominent perspective in the literature is a *societal impact* perspective, where the focus is on the ways in which self-regulation failure on the individual level contributes to certain developments in society. Typically, the structure of arguments within this perspective is such that the legitimacy of interventions on the individual level flows directly from the (assumed) problematic nature of the societal phenomenon to which they purportedly contribute. Consider this opening paragraph from Baumeister and Heatherton's (1996) seminal paper on self-regulation failure.

> Modern American society suffers from a broad range of problems that have self-regulation failure as a common core. Crime, teen pregnancy, alcoholism, drug addiction, venereal disease, educational underachievement, gambling, and domestic violence are among the social problems that revolve around the apparent inability of many individuals to discipline and control themselves. (Baumeister and Heatherton, 1996, p. 1)

Another example is the way in which self-regulation failure is said to be contributing to the obesity epidemic (e.g., Miller, Horodynski, Herb, Peterson, Contreras, Kaciroti, Staples-Watson, and Lumeng, 2012; De Ridder, De Vet, Stok, Adriaanse, and De Wit, 2013), which has been reported as a cause of rising health-care costs.[10]

In many texts about these subjects, the exact normative framework from which the societal phenomena are deemed problematic is left implicit. Typically, just as in the quote above, the phenomena that are discussed appeal strongly to intuition about what is "evidently

[10] For example, Finkelstein, Fiebelkorn, and Wang (2003) calculated that overweight- and obesity-attributable medical spending for the United States "accounted for 9.1 percent of total annual U.S. medical expenditures in 1998 and may have been as high as $78.5 billion ($92.6 billion in 2002 dollars)" (Finkelstein et al., 2003). In 2003, in the Netherlands, health care costs related to the effects of overweight amounted to nearly 1.2 billion euros (Van Baal, Heijink, Hoogenveen, and Polder, 2007).

problematic". The subsequent reasoning then generally follows the pattern that if the problematic nature of the societal phenomenon is accepted, and the idea that the phenomenon is (in part) caused by self-regulation failure is endorsed, then there are reasons for combatting the societal phenomenon through self-regulation interventions on the individual level.

Besides the societal impact perspective, there is a second perspective that is prominently featured in the literature, namely that of *individual health and well-being*. Here, the focus is on the ways in which people's self-regulation failures undermine their goals concerning health (e.g., their body-mass index, blood pressure or cholesterol levels) and well-being (e.g., their mood). One benefit of this perspective is that, by making the individual agent central, it takes away some of the reservations people might have about decisions that follow from the societal impact perspective, where the societal outcome can sometimes take precedence over the consequences for the individual (e.g., requiring individuals to increase their physical exercise with the aim of reducing health care costs).

From this individual health and well-being perspective more reasons for developing interventions that target self-regulation failure can be identified by examining empirical findings. While here is not the place for an exhaustive review of the literature (but see Baumeister et al. (1994)), it will be instructive to consider a number of key findings. For example, poor self-regulation has been associated with increased obesity risk (e.g., Fan and Jin, 2014) as well as an increased risk of attracting venereal diseases (e.g., Hernandez and Diclemente, 1992; Baumeister et al., 1994). Even increased morbidity and mortality rates have been linked to self-regulation failure as people fail to adhere to treatment programs (e.g., Dunbar-Jacob and Schlenk, 2000; Christensen, Moran, Wiebe, Ehlers, and Lawton, 2002) or procrastinate on important health behaviors such as making doctor's appointments (Sirois, Melia-Gordon, and Pychyl, 2003; Sirois, 2004). In addition, Kroese, De Ridder, Evers, and Adriaanse (2014) have recently suggested that poor self-regulation in the form of bedtime procrastination lies at the basis of much sleep deprivation in the general population (Kroese et al., 2014), which in turn has been linked to impairments in memory and concentration and is a major risk factor for developing depression, chronic hypertension, obesity and diabetes

(Harrison and Horne, 2000; Strine and Chapman, 2005; Buxton and Marcelli, 2010; Xiao, Arem, Moore, Hollenbeck, and Matthews, 2013).

Taken together, empirical findings such as the ones just described suggest that, from the individual health and well-being perspective as well, there are reasons for wanting to develop interventions aimed at reducing self-regulation failure. In addition, viewing the evidence from this perspective helps to understand the growing demand among consumers themselves for products that could support them in their self-regulation efforts.

In practice, intervention studies as well as commercial products are often promoted by referencing one or both of these two perspectives. For example, Boh et al. (2016) use the societal impact perspective to support a technological intervention to target obesity, by focusing on the obesity-associated health care costs (Boh et al., 2016). On the commercial side, the Philips DirectLife programme, which was geared towards helping individuals live a healthy lifestyle, was promoted by mentioning both the individual health benefits of the programme and the fact that the programme helps to reduce sick leave and healthcare costs (Philips, 2010).

In short, the reasons stemming from the individual health and well-being perspective as well as from the societal impact perspective are the main drivers for the development of self-regulation interventions. However, I want to emphasize a third perspective on self-regulation interventions, one that is sometimes overlooked outside of philosophy but that is very relevant to this topic as it brings into focus a different set of reasons that have less to do with (negative) consequences of self-regulation failure and more with the nature of failing in itself. For even if self-regulation failures have no significant negative consequences, there may still be reason to think that these failures are problematic from an agential point of view, i.e., from a perspective of *self-governing agency*.

As mentioned in the Introduction, I share the established position that self-governing agency is intrinsically valuable. By this I mean that there is value in having genuine opportunities for developing, maintaining and exercising capacities needed for considering for ourselves which values we hold dear, generating personally relevant ends, determining which of these ends we believe are worth striving for, reasoning about the means to those ends, and having the authority to

act in ways that allow us to attain those ends. While this is only an informal characterization—I will say more about how I conceptualize self-governing agency in Chapter 4—it is sufficient to get the point across that self-governing agency involves setting standards for oneself, as well as monitoring one's behavior and striving to make one's behavior meet these self-imposed standards. As such, how well we govern ourselves appears to stand in direct relation to how well we regulate ourselves. Indeed, I believe a certain degree of self-regulation success is a necessary condition for self-governing agency.

If this is right, then under the assumption that we place value in self-governing agency, there is cause to also consider self-regulation failure in relation to self-governance. After all, if self-governing agency involves being successful at self-regulation to a certain degree, then self-regulation failures can be considered a cause for concern, independent of their outcome. As a case in point, consider someone who

- intends to hand out three compliments each day to various people but never does,
- intends to avoid stepping on any edges of tiled pavement but frequently steps on them, and
- intends to be considerate with regard to sending birthday cards to friends and family but often cannot be bothered to send them.

This example is admittedly a little contrived, but that is because I want to highlight a scenario in which we may plausibly assume that there are no negative consequences of these failures in terms of either the person's health and well-being or society at large. Assuming this is the case for the three above-mentioned failures, I want to argue that we may still want to say—depending of course on the exact conception of self-governance that we subscribe to—that these failures are problematic because they suggest a shortcoming in the person's capacities to effectively determine one's own course in life.

Note that I am not advocating that people have to be without failure. Failures to self-regulate are, to a certain extent, perfectly compatible with self-governing agency, and perhaps even necessary, insofar that they help individuals to learn about themselves and their capacities for self-determination (e.g., what realistic goals are, how to make trade-offs between different courses of action, etc.). The point, rather, is to show that systematic failures of self-regulation can be indicative of a

kind of agency in which there is reduced self-governance or even no self-governance to speak of and that this is problematic irrespective of the consequences of one's behavior for one's health or well-being. If this indeed is the case, and we consider self-governing agency something to strive for, then we can see how this perspective might offer additional reasons for designing interventions that can support people in their efforts to regulate, and govern themselves.

With all of this now out in the open, it is easy to see the force of the different arguments in favor of developing interventions. At this point, however, I want to turn to the second part of my twofold aim for this section and stage an intervention of my own. For I want to emphasize how important it is to realize that even if we accept that one or more of these perspectives offer sound reasons for developing interventions, these reasons for intervening are still only *pro tanto* reasons that may be outweighed by other, more weighty reasons for refraining from intervening. Such reasons could, for example, stem from concerns about the extent of privacy people would have to give up in order for the interventions to be successful, or about the potential restriction of liberty the interventions would entail. I will return to issues such as these in Chapter 3. First, though, let us turn our attention to the type of self-regulation interventions that are currently being promoted so that we may have a better sense of the kinds of risks they bring with them.

Chapter 2

E-Coaching Systems

"[W]e should seek not only to elucidate the concepts we have, but aim to improve them in light of our legitimate purposes."

Sally Haslanger

The aim of this chapter is to gain conceptual clarity about the kinds of systems that we will be reflecting on in the chapters to come. Fully grasping these system's capabilities and their modes of influence is essential, because failing to do so makes one susceptible to making mistakes in the assessment of the possible risks involved with using these systems. That is, if one were to think of e-coaching systems as nothing more than convenient self-help tools, much like advanced calendaring systems with context-driven notifications, then it would be understandable if one were to downplay, or even outright dismiss concerns having to do with the ways in which sophisticated, adaptive systems form their own perspective on a user's health and behavior, and from that perspective give shape to user interactions that the user has not explicitly endorsed.

However, once we have in full view that e-coaching systems will involve a level of sophistication and independence that transforms the way we perceive them and interact with them, we can see more clearly the concerns that may arise with these advanced systems. In particular, it will help to understand in later chapters both why these kinds of systems can facilitate complacency in people's practical reasoning, and why this complacency can pose a threat to people's personal autonomy.

The task of attaining the sought conceptual clarity is made some-

what difficult by two factors. The first is the fact that these systems are only just entering the public sphere and are therefore not familiar to most people. This in itself has two implications. First, it means that understanding what these systems will be capable of will require some effort and imagination. To aid the imagination I will sketch and discuss a number of hypothetical systems and scenarios to make these systems as concrete as I possibly can at this point in time. Second, it means that a classification of these systems cannot be done in terms of their technical components (e.g., types of sensors, actuators, or algorithms) because these will most certainly change in the near future. Therefore, I will work towards a characterization of these systems in terms of their functional capabilities.

The second factor that sometimes stands in the way of a clear understanding of the sort of systems that this dissertation is concerned with, is that the term "e-coaching" is also used to describe the practice of coaching *through* technology. In this very broad sense, this means that if a human coach uses technology as a mode of communication (e.g., to get information about a coachee's behavior via Facebook or to give feedback via email or text message), this is considered e-coaching. Consequently it could be argued that the communication systems that human coaches and coachees use to communicate in this practice are types of "e-coaching systems". I want to establish here, at an early stage of the dissertation, that this is not the sense in which we should think about these systems in this dissertation. Rather, the term "e-coaching system" should be understood as referring to systems that are not just *facilitating* the coaching, but are actually *doing* the coaching. In other words, what we are engaged in is philosophical reflection on human-computer interaction rather than computer-mediated communication. Having this distinction in the background will hopefully help in attaining the conceptual clarity that this chapter aims to provide.

The chapter is structured as follows. First, in order to get a sense of what I mean with "a system doing the coaching", I will briefly elaborate on what I take coaching to be. Second, in Section 2.2, I will discuss three initial characterizations of "e-coaching systems" as found in recent literature. There, I will argue that these are broad and inclusive characterizations that do not sufficiently capture the level of sophistication and independence that is associated with a genuine process of coaching. Third, in Section 2.3, I will attempt to remedy this short-

coming in the literature by drawing on existing concepts from the respective fields of autonomous agent systems and behavior change support systems and proposing a list of eight features that I believe are necessary for systems to have in order to engage in a process that can be considered coaching. Then, from that list, I will derive a more narrowly construed definition of e-coaching systems, and discuss three key implications of the proposed definition. Finally, in Section 2.4 I will show how the definition brings to the foreground the impact that the sophistication of e-coaching systems will have on the socio-technical relationship that people maintain with these types of technologies.

2.1 Characterizing Coaching

In order to meaningfully define what an e-coaching systems is, we first need a rough approximation of what coaching is. Unfortunately, there are many definitions of coaching, with very little consensus about them (see for example Hayes and Kalmakis, 2007; Ives, 2008). Ives (2008) distinguishes no less than nine different approaches, ranging from humanist approaches (where the focus is on personal growth) to behaviorist approaches (where the focus is solely on changing behavior) and from cognitive approaches (where the focus is mainly on developing adaptive thoughts) to goal-oriented approaches (Ives, 2008). With the latter type of approach, coaching is "essentially about helping individuals regulate and direct their interpersonal and intrapersonal resources to better attain their goals" (Grant and Stober, 2006, p. 153).

With designers and developers of e-coaching systems, the goal-oriented approach appears to be the dominant strategy. While different approaches to offering support are certainly implemented—some systems put more emphasis on changing people's attitudes while others focus more on the behavioral outcomes—the support they offer is typically in the service of goal striving. Let us therefore examine what a goal-oriented coaching approach consists in. Ives writes:

> The primary method is assisting the client to identify and form well crafted goals and develop an effective action plan. The role of the coach is to stimulate ideas and action and

> to ensure that the goals are consistent with the client's main life values and interests, rather than working on helping the client to adjust her values [...] (Ives, 2008, p. 102)

In much the same vein, Hayes and Kalmakis (2007) paraphrase Stober as stating that "coaching is viewed as a customized, collaborative relationship that elicits the client's potential for self-awareness, for understanding the meaning of his or her unique situation, for visualizing change, and for making choices and plans to achieve a goal" (Hayes and Kalmakis, 2007, p. 556).

What both citations highlight is that coaching is an ongoing process between two parties who have a collaborative relationship that focuses on creating opportunities for improving self-understanding, increasing self-monitoring and supporting people's plan-making in order to improve goal striving. A good definition of e-coaching systems should take this into account. In what follows, I will review the current state of the literature on e-coaching systems to see how this relatively new term is being used, show that the current conceptualizations of e-coaching systems are broad and overly inclusive, and present my own, more narrowly construed definition to underline the level of sophistication that these systems have.

2.2 E-Coaching Systems in the Literature

In November 2014, the Dutch Rathenau Institute published an advisory report on the current developments concerning e-coaching systems (Kool, Timmer, and Van Est, 2014). Written for a broad audience, it discusses the trend observed in coaching practices towards the digitalization of both coachees and coaches and gives an assessment of the societal impact that these developments may have. In the report, the notion of "e-coaching system" is understood in a very broad sense so that it covers a wide variety of systems.

The authors characterize e-coaching systems in terms of three processes: collecting data, analyzing data and determining a coaching strategy, and giving persuasive and motivating feedback (Kool et al., 2014, p. 16). While these conditions might seem quite stringent on first glance, they are actually quite lenient. So much so, in fact, that the class of systems these conditions carve out include systems of which it is not

clear in what sense they are taking on the role of a coach. To give an illustration, consider the "Mimo Baby Monitor" (`http://mimobaby.com`) that the authors themselves include as an example of an e-coaching system (Kool et al., 2014, p. 195). It is true that this baby monitoring system is technologically advanced in comparison to other baby monitors, in that it reports on a variety of measurements such as body position and skin temperature, and that it offers parents the possibility to configure it to send notifications under certain conditions. Surely such a system may be part of a coaching practice (e.g., coaching first-time parents to take proper care of their infant child) but it is far from evident that the system itself is doing any coaching.

In an earlier paper, Warner (2012) characterized e-coaching systems as *pedagogical agents* that provide "questions to coachees and responses based on coachees' entries or selections" (Warner, 2012, p. 24). The positive of this characterization is that it highlights the dialogical nature of coaching. The downside however is that the characterization seems at once too broad and too narrow. It can be considered too narrow because pedagogical agents are typically defined as having life-like animated interfaces (see e.g., Gulz, 2004; Mayer and DaPra, 2012), whereas with e-coaching, it is perfectly possible to be coached solely by voice, text or through other interfaces (see for example Beun et al., 2016).

At the same time, Warner's characterization appears to be too broad in that the level of sophistication of the system is specified very minimally in terms of responding appropriately according to input choices made by the user. If this definition is taken at face value, a simple keyword-based algorithm like that used in Weizenbaum's (1966) "Rogerian therapist" ELIZA (Weizenbaum, 1966) could plausibly power a pedagogical agent that under Warner's definition would be considered an e-coaching system. Again, however, it not clear that such a system should qualify, as it is not evident that the system is engaged in a process of coaching.

In Warner's defense, it can be argued that additional criteria that determine the system's minimal level of sophistication is packed into the notion of pedagogical agent. For when computer scientists talk of "agent systems"—or simply "agents"—they are, roughly, referring to computerized systems that, at a minimum, are *embedded in an environment* in which they can *sense* and *operate* and to which they are

reactive (e.g., Brustoloni, 1991; Wooldridge and Jennings, 1995; Franklin and Graesser, 1997; Russell and Norvig, 2003). Beyond these core elements, however, there is no consensus about the constitutive criteria for agent systems. Wooldridge and Jennings (1995) for example see *autonomy*—in this context understood as operating without constant human guidance—as a constitutive criterion for agent systems (Wooldridge and Jennings, 1995, p. 2), while other authors understand the "autonomy" feature as carving out a particular subclass of agent systems (e.g., Russell and Norvig, 2003). Likewise, a number of additional criteria have been proposed that may or may not be considered constitutive to agent systems, most notably *social ability* and *proactiveness* (e.g., Wooldridge and Jennings, 1995, p. 2). The former is the ability to communicate with other agent systems—for example in *ambient environments* (Aarts and De Ruyter, 2009)—and with human beings through some kind of language. The latter is the idea that "agents do not simply act in response to their environment, [but] are able to exhibit goal-directed behaviour by taking the initiative" (Wooldridge and Jennings, 1995, p. 2). In addition, it has been suggested that agents, and pedagogical agents specifically, ought to be able to learn in order to be *adaptive* (e.g., Giraffa and Viccari, 1998). As all of these features are highly relevant for coaches, it could be argued that, aside the first point about the animated avatars, Warner's characterization is in fact apt.

However, recall that the concern was that the characterization was overly inclusive, not that it could not be stretched to cover the whole spectrum of e-coaching systems. Though a charitable reading of Warner's characterization certainly is less inclusive than the one by Kool et al. (2014), it remains vague about which of the features just mentioned are strictly required for a system to count as an e-coaching system. Consequently, the door is left open for interpretations in which these features do not play a role.

To be clear, I do not mean to rule out that e-coaching systems can have varying levels of sophistication; in fact, I contend that they can and do. The point is rather that the minimum requirements for classifying as an e-coaching system should be more strict—and, in any case, more explicit—in order to be able to say something meaningful about the entire class of systems.

Finally, Van Wissen (2014) has characterized e-coaching systems as

behavior change support systems (BCSSs) (Van Wissen, 2014, p. 5). This term was coined by Oinas-Kukkonen in 2010, and defined as an "information system designed to form, alter or reinforce attitudes, behaviors or an act of complying [...]" (Oinas-Kukkonen, 2010, p. 6). In this formulation, BCSSs are introduced on a par with *persuasive systems*: "interactive computing systems designed to change people's attitudes and behaviors" (Fogg, 2003, p. 1). In later work, Oinas-Kukkonen (2013) added that a "special characteristic of BCSSs is that they request [...] emphasis on positive user experience and stickiness to motivate users to engage with them regularly over an extended period of time" (Oinas-Kukkonen, 2013). This quote highlights that there will be repeated interactions between the user and the BCSS and that the outcome of the interaction—where the outcome will often be the user's behavior—will play a role in the feedback that the system gives to the user. Still, adding these characteristics does little to narrow down the class of systems that qualify because the technical requirements remain very minimal. In fact, as Oinas-Kukkonen understands the term, it includes both human-computer interaction and computer-mediated communication (Oinas-Kukkonen, 2013, p. 1227). Moreover, on the human-computer interaction side of things, Oinas-Kukkonen himself allows fairly simplistic systems such "an interactive picture frame for adopting better sitting habits while working at the computer (Obermair, Reitberger, Meschtscherjakov, Lankes, and Tscheligi, 2008)" (Oinas-Kukkonen, 2010, p. 5) to count as a BCSS.

On the face of it, then, it would seem that Van Wissen would allow this same interactive picture frame to qualify as an e-coaching system. However, from her other remarks in Van Wissen (2014), it becomes clear that she is solely concerned with human-computer interaction *and* that she presupposes more complexity in BCSSs than is inherent in the definition of those systems. For example, in Klein et al. (2014), Van Wissen and her colleagues focus on "accessibility, adaptability and interactiveness" (Klein et al., 2014, p. 138). Moreover, they make an explicit distinction between *simple*, *tailored* and *model-driven* e-coaching systems (Klein et al., 2014, p. 139). The first relates to generic interventions, the second to tailored interventions, and the third to systems that rely on theory-driven models of behavior change to determine the root causes of people's non-compliance. What this highlights is that the authors are distinctly aware of the different levels

of sophistication that behavior change support systems can have. The open question that concerns us here, however, is whether all of these systems should classify as e-coaching systems.

Van Wissen seems to suggest that they should. She characterizes e-coaching systems along three axes of a three-dimensional space. The axes are the "complexity of persuasive techniques, the use of artificial intelligence techniques and whether they utilize user models that have a solid theoretical foundation" (Van Wissen, 2014, p. 261). This idea is very helpful for understanding on a high level the different aspects that are related to e-coaching systems, but Van Wissen fails to be specific about the lower boundaries of the concept "e-coaching system". Though she is adamant that researchers should strive for finding "the e-coaching sweet spot"—where the different dimensions are all represented to some sufficient level—she does seem to allow that systems that have few persuasive techniques, little to none artificial intelligence techniques and are not model-driven *can* be called e-coaching systems.

The advantage of casting a wide conceptual net in each of these cases is of course that it allowed the authors to discuss a wide range of technological developments and to discuss various products already on the marketplace.[1] The downside of such a broad conception, however, is that it allows for simple systems with which we are familiarized in the present to shape our thinking about interactions with technologies of the future. This may lead to a systematic underestimation of the risks involved with using e-coaching systems, which in turn may lead to missed opportunities to proactively make informed policy decisions that take into account the level of sophistication and independence that these systems will have.

To remedy this situation, I will stipulate eight features that I believe are *necessary requirements* for e-coaching systems, and from those features derive a more narrowly construed definition. Then, I will show that with that definition in place, it becomes much clearer that the sophistication of e-coaching systems changes the socio-technical relationship that people maintain with these types of technologies.

[1]This approach is understandable particularly for Kool et al. (2014), considering that their report is specifically concerned with the trend *towards* autonomous e-coaching systems.

2.3 Defining E-Coaching Systems

As discussed in Section 2.1, if systems are to be engaged in coaching, it is key that they are able to create and maintain customized, collaborative relationships in which coachees are supported in understanding their situation and in making effective plans for changing their behavior or attitudes in accordance to their own view on how to live their lives. With this in mind, I propose that, minimally, these systems should be able to do the following:

1. Engage in an ongoing conversation with the user. This conversation is crucial for establishing and maintaining a collaborative relationship between user and system. For this, the system will need to implement *social ability* (see again Section 2.2).

2. Be perceived as *credible*, i.e., as having expertise and being trustworthy (see Flanagin and J., 2008, p. 8). This credibility is crucial since coaching requires repeated interactions between user and system in which the user should be given reason to give serious consideration to the information that the system conveys.

3. Be *context-aware*. In order to stimulate ideas and action, and to assess whether a person's goals are consistent with that person's life values, the system will need to be in some relevant sense aware of the context in which particular behaviors take place and particular decisions are made. For more on the importance of context for e-coaching systems, see Van Wissen, Kamphorst, and Van Eijk (2013).

4. Ask questions, give feedback and offer advice that is *tailored* (Fogg, 2003) to the individual user. That is, though the coaching may be based on underlying, generic principles, the interactions should be sufficiently specific to the preferences and behaviors of the individual to make the relationship customized. In order to perform the tailoring, the system will need *learning abilities* to build up and maintain a *personalized user model* (cf. Klein et al., 2014; Van Wissen, 2014).

5. *Interface* with (heterogeneous) data streams (e.g., direct user input, but potentially also measurements of physical activities, mood

self-reports, sleeping patterns, etc.). Only if the system has sufficient information will it be able to ask the appropriate questions and make fitting recommendations.

6. Be *proactive* (Wooldridge and Jennings, 1995, see p. 2) in its interactions with the user. This may mean that it may initiate interactions (e.g., by sending push messages to one's smartphone to warn the user at suspected moments of weakness), but more importantly, it means that the system can co-determine what the conversation is about, and that it can stimulate action or reflection (e.g., invite a user to reflect on a his or her commitment to a particular goal).

7. Monitor one's behavior change progress. If the system is to be successful in coaching people to change their behavior, it needs to have some notion of what a behavior change trajectory looks like. For this, it needs to operate on some type of *model of behavior change* (cf. the COM-B model (Michie, Ashford, Sniehotta, Dombrowski, Bishop, and French, 2011) and the COMBI model (Klein, Mogles, and Van Wissen (2013); see also Kamphorst et al. (2014a)).

8. Guide the user in a process of future-directed intention formation, also known as *planning*. As seen in the definitions of goal-directed coaching, planning support is key for setting users up for behavior change success.

These eight features give shape to a class of systems that go beyond the mere persuasive presentation of collected data, and also beyond straightforward triggers, cues and environmental scaffolding for behavior change. They make explicit that e-coaching systems are adaptive over time to a user's changing preferences and behavioral patterns, that they are in a relevant sense context-aware, that they can predict what a user is about to do, or how the user will respond to a certain interaction, and that they are able to engage in a conversation with the user in which the system can take initiative and co-determine what the conversation is about. These ingredients, I believe, form the basis for systems that can engage with the user in an interactive process that can reasonably be considered goal-oriented coaching in line with the definitions discussed in Section 2.1. Taking these elements together, I propose the following definition of e-coaching systems.

> *E-Coaching System.* An e-coaching system is a set of computerized components that constitutes an artificial entity that can observe, reason about, learn from and predict a user's behaviors, in context and over time, and that engages proactively in an ongoing collaborative conversation with the user in order to aid planning and promote effective goal striving through the use of persuasive techniques.

Defining e-coaching systems in this way has a number of implications. First, and most obviously, this more narrowly construed definition helps to distinguish e-coaching systems from a range of other self-regulation support structures. For example, it will be evident that persuasive photo frames and smart baby monitoring systems do not qualify as e-coaching systems. That is not to say that such systems—I propose to call them *self-regulation facilitators*—cannot play a supporting role in changing people's behaviors or attitudes, or that they cannot be included in a coaching practice. However, it does mean that such systems should be excluded from discussions that are strictly about policies concerned with, and the ethics of, computerized e-coaches.

Second, the definition helps to counter a contention that often comes up in relation to e-coaching technologies, namely that such systems can be manipulative by bypassing people's rational capacities. What the definition makes clear, is that though e-coaching systems may use persuasive techniques, they cannot bypass people's rational capacities altogether, by virtue of the ongoing dialogue between the system and the user. Of course, the user need not be aware of all persuasive techniques being used, and therefore some risks of manipulation remain—parties may have incentives to subtly steer the coaching conversation towards the use of certain brand products, for example—but having people understand that they are engaged with a goal-directed system does ensure that people have an opportunity to arm themselves in what Dennett has called "the arms race of techniques of persuasion" (Dennett, 2014, p. 583).

Finally, related to the previous point, the definition brings to the foreground that e-coaching systems operate with a level of sophistication that causes a subtle shift in the dynamics of the *socio-technical relationship* that people develop and maintain with technologies (cf. Alberdi, Strigini, Povyakalo, and Ayton, 2009). One way in which this shift is reflected, is the increased complexity of interactions that people

will have with these systems. As systems improve their ability to process natural language, for example, the interactions will naturally evolve from fairly simplistic "option picker" interactions to forms of dialogue.

But the shift in dynamics extends beyond form alone. Just consider the contrast between interacting with a more "tool-like" system such as a digital agenda and an e-coaching system. With an agenda, the user is fully in charge of when interactions will take place, and the user knows what the content of those interactions is going to be (e.g., a reminder for an upcoming meeting). With an e-coaching system, on the other hand, the timing of the interactions will be (co-)determined by the system (e.g., based on a weighed assessment of a user's preferences and the predicted relevance of an interaction in the user's current context), and the content may contain information that is partially or wholly new to the user.

While these differences may seem insignificant from a usability perspective, and to users may feel like a natural progression of how we use tools, the uncertainty that is introduced into the interactions actually marks a crossing of an important threshold of system complexity that carries normative implications for how we, as agents using these technologies, should relate to these systems. I will return to these implications in Chapter 5. First, however, I will conclude this chapter with an introduction to the theses of the *extended mind* and the *extended will*, in order to have a rich vocabulary with which to characterize the social-technical relationship between user and e-coaching system.

2.4 The Socio-Technical Relationship

In recent years, much work has been done in philosophy and cognitive science to enrich the vocabulary with which the socio-technical relationship that people develop with various technologies can be described. This body of work pertaining to embodied, embedded, extended and enacted cognition is sometimes referred to as "4E cognition" (Menary, 2010b). Of particular interest to discussions about e-coaching systems is that theorists in this domain have been working out the implications of taking seriously the idea that agents' cognitive and volitional processes are not restricted to agents' bodies, but that they are embedded in an environment, and sometimes tied so closely

to parts of the environment that it is apt to speak of an *extended* cognitive or volitional system (e.g., Clark and Chalmers, 1998; Menary, 2006; Rowlands, 2009; Heath and Anderson, 2010; Wheeler, 2011; Clark, 2011; Aydin, 2013; Gallagher, 2013).[2]

Proponents of "extendedness theses" see a major benefit in conceptualizing the mind as extending beyond the boundaries of skin and skull because doing so provides a natural way of describing the human capacity for taking something from the environment—at the moment of writing, a smartphone is an apt example—and integrating it seamlessly into our being. As Clark put it, "external tools [can] become *transparent in use* so that our intentions 'flow through' the tools to alter the world, that we feel as if we directly control the limbs, or tools, in question, that we begin to feel as if they are a part of us" (my emphasis; Clark, 2003, p. 123). Moreover, by taking the extendedness theses seriously, we can do justice to the role that environmental engineering plays in our lives (see again Section 1.2.1).

The theses of *extended cognition* and *extended will* both give expression to this line of thought by holding that under certain conditions of *functional parity*, "technological equipment may count [...] as constitutive parts of one's cognitive system" (Carter and Palermos, 2016, p. 546) or of one's volitional system (Heath and Anderson, 2010). This notion of functional parity was first described by Clark and Chalmers in their seminal 1998 paper "The extended mind". In that paper, Clark and Chalmers voiced their dissatisfaction with the common assumption that the mind is located inside one's head, and that it stops with "the boundaries of skin and skull" (Clark and Chalmers, 1998, p. 7). Driven by the observation that the environment often plays crucial roles in cognitive processes, Clark and Chalmers argued for a position they called *active externalism*, which maintains that *coupled systems* of human organisms together with external entities could under certain conditions be considered a cognitive system in their own right.

This forming of coupled systems, they argued, could happen if the following necessary principle would be met.

[2]Some authors, such as Aydin and Gallagher, go as far as to object to the notion of extendedness, because it reaffirms the inside/outside distinction that they would like to see abolished. This move follows quite naturally from their liberal views on cognition and the mind, but for me, the notion of extendedness signals that certain core properties of agents, by contingent empirical fact, do lie within the skin-skull boundaries. Therefore, I will continue to use this terminology.

> *Parity Principle.* If, as we confront some task, a part of the world functions as a process which, *were it to go on in the head*, we would have no hesitation in accepting as part of the cognitive process, then that part of the world *is* [...] part of the cognitive process. (Their emphasis; Clark and Chalmers, 1998, p. 8)

Moreover, in order to resist "an unacceptable proliferation of systems (many of them extremely short-lived)" (Rupert, 2004, p. 396), Clark and Chalmers proposed four additional criteria—sometimes referred to as the "trust and glue" criteria—that together with the parity principle would be sufficient for extended cognition[3]:

> *Reliable availability.* The external resource has to be "reliably available and typically invoked" (Clark, 2011, p. 79);
>
> *Automatic endorsement.* The information obtained from the external resource has to be more-or-less automatically endorsed by the agent, in that the information "should not usually be subject to critical scrutiny" (Clark, 2011, p. 79);
>
> *Easy accessibility.* The information in the external resource should be "easily accessible as and when required" (Clark, 2011, p. 79);
>
> *Prior endorsement.*[4] The information contained in the external resource has to be consciously endorsed by the agent in the past and should be there "as a consequence of the endorsement" (Clark, 2011, p. 79).

The prototypical example used to illustrate the idea of extended cognition concerns a person named Otto, who suffers from Alzheimer's disease, and for that reason relies on a notebook to help him structure his life. He carries the notebook with him always, writes down every bit of relevant information that he learns, and looks up information in the notebook when he needs it. In other words, as Clark and

[3] Note that, contra my reading, Aizawa takes these "trust and glue" criteria as having been suggested as alternative criteria to the parity principle and supposes that proponents of "trust and glue" believe that these criteria can be sufficient for extended cognition without the parity principle also holding true (Aizawa, 2018).

[4] Already in the original 1998 paper Clark and Chalmers expressed their uncertainty about this prior endorsement criterion, admitting that it might be overly restrictive (see Clark and Chalmers (1998, p. 17) and also Clark (2011, p. 80)). For that reason, some authors choose to omit the criterion, but I am keeping it for the sake of completeness.

Chalmers describe the case, Otto's notebook is reliably available and easily accessible, its contents is provided by Otto himself and the information contained in the notebook is automatically endorsed by Otto whenever he consults it. Moreover, Otto's notebook *functionally* does what biological memory does for others, namely allowing Otto to store and retrieve information. If Otto needs to recall the address of the Museum of Modern Arts, he consults the notebook, whereas others would access their biological memory.

Clearly, there are many differences between memory as instantiated in the brain, and Otto's notebook—e.g., portability, risk of damaging, learning-time and access-time, etc.; see Adams and Aizawa (2001); Rupert (2004); Sutton (2010)—but those are not the differences that matter, according to Clark and Chalmers. Otto and his notebook, they claim, together achieve the same feat as others who solely rely on their internal, biological memory to recall the address. I will not go into the details of the argument (but see Clark (2011)), but the upshot is that, because the notebook functions as (a crude form of) memory, and because we have no hesitation accepting memory as part of the cognitive process, we should allow Otto's notebook to be part of the cognitive system as well. In other words, on the basis of functional parity, the coupled system of Otto plus notebook should be viewed as a cognitive system in its own right. Otto's cognition, and, by extension, his mind, are not "just in the head," but extended out into the environment.

Continuing this line of thought, Heath and Anderson (2010) have suggested that besides cognition, "the will" is also better understood as being "transcranial" or "extracranial" (Heath and Anderson, 2010). As an illustration, consider how people set up calendar app notifications as triggers to get themselves to meetings they intended to be present at. According to the extended will thesis, engineered parts of the environment such as these could, under the right circumstances, be considered part of one's extended will on the basis of the parity principle since we would have no hesitation to consider the recall of an intention as part of one's will, were it to occur inside the head.

Of note is that Heath and Anderson do not themselves explicitly engage with Clark and Chalmers's proposed criteria for extendedness, but instead conceptualize the external resources that qualify as being part of an extended volitional system as *prosthetics* rather than "mere"

aids or tools. Contained in this articulation certainly appears to be an implicit acceptance of a reliability criterion, but it is less clear to what extent the authors subscribe to Clark and Chalmers's criteria about the endorsement of information (i.e. the automatic endorsement criterion and the prior endorsement criterion). It stands to reason, however, that (variants of) those criteria would be useful in distinguishing between external resources that do and do not become part of one's extended volitional system, just as the criteria aim to do for extended cognition. For example, in the case of the calendar app notifications, the information that is provided by the app will have been endorsed by the agent in the past when he entered the appointments in his calendar and enabled the notifications (prior endorsement), and the content of the notifications will be endorsed more-or-less automatically by the agent when the notifications are received (automatic endorsement). Of course, the agent may reconsider going to a particular meeting after receiving a notification about it, but he will not normally subject the notification itself to critical scrutiny in the sense of doubting the content or the timing of the notification.

In contrast, it would seem apt to say that the person who has enabled notifications for a shared, work-related calendar but who typically ignores the notifications, or at least scrutinizes them critically for relevance, is using the notifications as a source of information, without them becoming part of his extended volitional system. To put it in one of Clark's more recent phrases used to describe "the right kind of coupling," there is no "dense integration" in this scenario between the agent and the external resource.

It is possible, however, that Heath and Anderson perhaps would want to be more lenient or liberal with regard to the criteria for extendedness. If that were the case, they would not be alone. For example, in a treatment on socially extended cognition, Gallagher has argued that Clark and Chalmers's "trust and glue" criteria are simply "wrong-headed" (Gallagher, 2013, p. 5). Other authors have even moved away from the functional parity principle and have approached extendedness through a principle of complementarity instead (for discussion, see Sutton, 2010; Heersmink, 2015; Aizawa, 2018). With this approach, the focus is less on extended resources implementing a function that could also be realized biologically inside the skin-skull boundaries (e.g., Otto's notebook fulfilling the same functional role as biological

memory), and more on the external resources playing a role in the overall cognitive system that complements the agent's internal cognitive processes (e.g., a calculator performing arithmetic operations as part of an agent's efforts to solve a mathematical problem).

Finally, in recent years, attempts have been made to analyze integration (and thereby extendedness) as a matter of degrees along various axes in a multi-dimensional space, with dimensions including the nature of the external resource (natural, technological or sociocultural), the durability and reliability of the overall system, the level of trust involved, the level of procedural transparency, and more (Wilson and Clark, 2009; Sterelny, 2010; Sutton, Harris, Keil, and Barnier, 2010; Menary, 2010a; Heersmink, 2015).

As will be evident from the ongoing work in this domain, working out the exact criteria for extendedness is beyond the scope of this section. The disputes in the literature concern deep metaphysical questions about the boundaries of cognition, volition, and minds that I will not attempt to adjudicate here (but see Adams and Aizawa (2001); Rupert (2004); Menary (2006); Wheeler (2010); Clark (2011); Heersmink (2017); Aizawa (2018)). Rather, what I want to draw attention to, is how thinking about the socio-technical relationship between agents and external technological resources in terms of extendedness, "seamless integration", "transparency in use", etc. brings front and center the kind of close-knit coupling that can happen with technology, and the trust that is required for this kind of relationship.

Seeing the relevance of trust sharply in turn helps to lay bare a tension between trust and vigilance that is especially relevant with regard to e-coaching systems. On the one hand, as I have detailed in this chapter, e-coaching systems are designed to be adaptive over time to a user's changing preferences and behavioral patterns, to be context-aware, and to predict what a user is about to do or how the user will respond to a certain situation. In this sense, they are equipped with capabilities that facilitate trustworthiness and reliance and make them good candidates for seamlessly integrating into one's life, and, perhaps even, one's mind (cf. how Carter, Clark, and Palermos (2018) hold that a "super-app" that is adaptive to one's context and performs various "smart," automatic information-transforming functions can be part of one's extended mind). On the other hand, however, in light of their reasoning and learning capabilities and their level of independence,

it may be questioned whether the kind of close coupling between agents and e-coaching systems that this yields, especially when there is automatic endorsement of the information that the e-coaching system provides, is really desirable. Since e-coaching systems, contra self-regulation facilitators, will not simply repeat back to the user what he or she has put it, but instead will offer transformed or wholly new information in persuasive ways, it would seem pertinent that the agent remains vigilant with regard to the suggestions made by the e-coaching system. This is especially so when e-coaching systems are employed to promote certain patterns of behaviors and to help undo other patterns, which may well lead to suggestions that are in service of those goals but are not perfectly aligned with what the agent is comfortable doing or would want to do if he were to critically reflect on his situation.

It is this tension between trust and vigilance that is brought out by theorizing about extendedness, and it is this tension that will play a central role in the remaining chapters of the dissertation, where I will argue that users of e-coaching systems should be careful not to become complacent with regard to processes related to practical reasoning. In the next chapter, I will first give an overview of the various ethical concerns that can arise with e-coaching systems in general, but will subsequently give a first approximation of this complacency concern.

Chapter 3

Ethical Concerns of E-Coaching: A Preliminary Inventory

"All over the place, from the popular culture to the propaganda system, there is constant pressure to make people feel that they are helpless, that the only role they can have is to ratify decisions and to consume."

Noam Chomsky

Now that we have more conceptual clarity about what constitutes e-coaching systems, this chapter will be dedicated to making a preliminary inventory of types of ethical concerns that arise with the emergence of these systems specifically (as opposed to self-regulation support technologies in general). This inventory is meant to serve a dual function. First, it is meant as an initial framework for guiding and structuring future discussions around the ethics of e-coaching systems that go beyond the specific focus of this dissertation. The aim is to make a contribution in this regard by separating the various concerns that e-coaching systems may give rise to, even if I will not be able to give a satisfactory treatment of all of them here. This brings us directly to the second intended function of the inventory, which is to make clear which concerns are at the heart of this dissertation. The inventory will thus help to situate my main argument about complacency in

relation to the wider ethical landscape.

In what follows I will discuss six sets of related but distinct concerns, which I have classified into three main categories that make up the sections of this chapter. The first category addresses concerns about social justice, broadly construed. Here I will survey questions related to how a widespread adoption of e-coaching systems may affect individual liberty, access to resources, cultural criticism, distribution of reputation and other fairness considerations. The second category relates to concerns about (intentional) infringements of rights of autonomy. Under this heading I will discuss concerns about coercion and manipulation, as well as worries about infringements of rights such as the right to privacy. Finally, the third category comprises a single concern having to do with the potential effects on people's exercise of self-governance stemming from the interplay between e-coaching technologies and their users. Central to this category will be the idea that valuing personal autonomy entails a responsibility for users to be vigilant to a certain degree with regard to the suggestions they receive.

Before we begin, let me make one final remark about the scope of this chapter. As will be clear after reading Chapter 2, e-coaching systems work primarily on a psychological level. However, it is imaginable that there could also be e-coaching systems that in addition to giving advice and feedback, also control parts of our bodily functionings more directly. Think for example of a chip that is implanted in the brain and that is hooked up wirelessly to the e-coaching system so that it can control the brain's reward system by controlling the flow of dopamine. Interventions on this level, which can be likened better to doping than to coaching, raise different ethical concerns. Those concerns are equally important, but outside the scope of this dissertation altogether.

3.1 Concerns about Social Justice

Due to its broad scope, the term social justice is not easily defined (Reisch, 2002). It is, however, typically accepted that social justice concerns the fair distribution of wealth, welfare, opportunities and privileges in society. So understood, social justice is tightly connected to discussions about individual rights and the negative and positive duties that flow from these rights for governments, social institutions

and individuals themselves. The widespread adoption of e-coaching systems potentially affects a wide variety of social justice considerations. It raises a number of pressing questions, for example about who will have access to e-coaching technologies, the extent to which the use of e-coaching systems will give its users a competitive advantage over those who do not use such systems, and the way in which this could generate a culture of competitive self-management. In this section I will survey three sets of such concerns, beginning with concerns about equal access. Subsequently, I will move on to concerns about liberty restrictions and, finally, to concerns about diminishing welfare through potential loss of self-esteem, reputation and social status.

Concerns about Equal Access

In arguing in favor of e-coaching systems, and in line with some of the reasoning strategies discussed in Chapter 1, proponents of e-coaching systems have a tendency to assume that the introduction of e-coaching systems will only make coaching more readily available to all. This type of reasoning typically runs along the following lines: given the significant costs associated with human-to-human coaching, coaching is currently mostly reserved for people with moderate to high incomes. Since e-coaching systems are projected to become much cheaper alternatives to human coaches, e-coaching systems could thus potentially reach a larger population than is currently the case with human-to-human coaching.

While this reasoning in itself is not invalid, it is important not to be drawn to the further, faulty conclusion that improved access to coaching actually implies more equal access. For even if e-coaching systems do indeed reach a larger audience, the products (including sensor systems, networks, smartphones, etc.) will in all likelihood not be free. Therefore it remains a real possibility that current societal inequalities may lead to unequal and unfair access to e-coaching systems, which in turn might aggravate the existing inequalities.

One hypothetical scenario of how this might happen—based on current market strategies—involves a proliferation of offerings on the market, ranging in price, but certainly also in quality. A parallel development can be seen in hardware products, where the more affluent have access to higher-quality smartphones and tracking devices. Big companies with big budgets for research and development can invest

more in validating their systems experimentally, as well as in making sure that they are keeping up with the latest technological advancements. However, investments like these will ultimately have to be financed, and typically this will be done by offering entry-level products and services for the masses to ensure significant market shares, and high-end products and services for a much more select group. Assuming that a similar pattern will emerge with the introduction of e-coaching systems, there is a risk that there will be a gap between those who can and those who cannot afford e-coaching systems at all, and also between those who can afford to buy evidence-based, quality e-coaching systems and those who can only afford entry-level models.

What makes this matter particularly pressing is that the net result of the inequality may have more far-reaching consequences than one group of individuals just having more expensive equipment, tools or gadgets than others. For if we suppose that the high-end systems deliver the goods and provide superior support to people's self-regulation, we can speculate that the improved self-regulation that ensues will deliver a competitive advantage to the affluent over people using lesser e-coaching systems or no e-coaching systems at all. If this is the case, then we can expect the gap between the privileged and the unprivileged to increase even further because of it, for example if people with access to high-quality e-coaching systems will be considered more attractive employees and will be more likely to be hired into high-salary positions.

This bleak scenario highlights a clear political issue that deserves careful consideration. One potential mitigating strategy has already been suggested in the literature, namely to allow insurance companies to distribute e-coaching systems of a certain baseline quality free of charge as part of their health insurance packages. However, as we will see in the following section, this approach raises another set of ethical concerns.

Concerns about Liberty Restriction

In many domains, public policies have already been implemented to encourage particular "desirable" behaviors by offering incentives to citizens. For example, in a number of Western countries there are monetary incentives for reducing the amount of trash collected (Thøgersen, J., 2003). Similarly, insurance companies are offering lower insurance

rates for individuals who can demonstrate that they are reducing their unhealthy habits such as alcohol consumption, tobacco use, poor diets and sedentary lifestyles. These incentive schemes are often promoted by calling attention to the social impact perspective or the individual health and well-being perspective as mentioned in Section 1.3.

Currently, the existing incentive programs are offered on an opt-in basis, and are fairly limited in their scope, in part because insurance companies at present do not have the means to accurately monitor compliance of people's various aspects of self-regulation. With the widespread adoption of e-coaching systems, however, and the increasing improvements made to measurement instruments, infrastructure, and data analysis techniques that go hand in hand with the development of e-coaching systems, this might all change in the near future. Technically, it will become much easier to extensively monitor people's behavior and compliance to their agreements with the insurance companies (e.g., not to smoke, or to exercise twice a week). This development will make it attractive for insurance companies to provide ever more fine-grained options for individuals to limit their insurance costs in exchange for information, and to nudge individuals to give up personal information in exchange for significant discounts on their insurance premiums (see for example Kool et al., 2014, pp. 48). The risk here is that, insofar as countries allow for such programs to be implemented, this development might lead to a situation in which people, in practice, will be unable to opt out of such incentives programs. As health-care costs (and insurance premiums) continue to rise, the less affluent citizens in countries with insurance-based health care systems may thus find themselves under increased pressure to choose one of these "restrictive-conditions" insurance policies, simply because they cannot afford to do otherwise. It is in this sense that the widespread adoption of e-coaching systems through the intervention of insurance companies could diminish those people's liberty to decline the use of e-coaching systems.

Given the value typically ascribed to liberty, this concern calls for a thorough and nuanced discourse in which this worry is addressed in relation to other liberty considerations, such as the one that increased insurance options for those making lifestyle choices already in line with what insurance companies find desirable, might be considered an expansion of their liberty. Moreover, given the interest that insurance

companies have already expressed in selling "interactive insurances" where data is traded for discounts (cf. John Hancock's interactive life insurances (e.g., Barlyn, 2018); see also Martani, Shaw, and Elger (2019)), this concern also calls for being placed on the political agenda with some urgency, together with more general questions about the legitimacy of such approaches to personalized insurance premiums.

Notice, however, that the kind of worry just discussed about liberty restriction is not strictly dependent on the role that insurance companies decide to play in the uptake of e-coaching technologies. For similar issues may arise also when a majority of individuals in the community is increasingly benefiting from e-coaching and gaining a competitive advantage, and others could come to be expected to keep up and meet rising expectations. A similar process can already be observed with smartphones, with more and more people being expected to carry a smartphone with them at all times, and to have enabled certain applications (e.g., Facebook's WhatsApp). The development of e-coaching systems could push this process even further if carrying smartphones would not only be associated with increased connectivity but also with increased self-regulation.

As mentioned at the outset of the chapter, my aim is not to analyze or attempt to resolve issues such as these here, but rather to call attention to them and placing them on the agenda for future discussions beyond this dissertation. For the purpose of creating the inventory that this chapter aims to establish, I want to focus now on another set of concerns, one having to do with the implications of using e-coaching systems for how people's technology-supported actions are perceived; in particular, the potential effects that shifts in action evaluation may have on people's welfare by negatively affecting self-esteem, reputation and social status.

Concerns about Self-Esteem, Reputation and Social Status

Thus far, there has been an implicit assumption in the discussions that successfully being supported by e-coaching systems in one's self-regulation will, in every respect, be beneficial to the person exhibiting more self-regulation as a result of being supported—potentially even leading to (potentially unfair) competitive advantages over those who do not use such systems. In Chapter 1 we have seen that there may be some intuitive plausibility for this assumption given the potentially

negative consequences of self-regulation failure, but the assumption can also be challenged. For what if being technologically supported in one's self-regulation will negatively affect how one's self-regulation success is perceived, either by oneself or by others? It is this line of thinking we turn to now.

First, it may be that with increased use of e-coaching systems in a wide variety of domains, individuals will less and less feel that they themselves, as agents, can take credit and feel a sense of accomplishment for their successful self-regulation. For example, just as people usually do not take credit for remembering all the telephone numbers stored in their mobile phones, so too could it become odd to feel accomplishment for regulating one's medications in relation to one's food intake throughout the day if one was in fact informed of the optimal timing by an e-coaching system. Yet this sense of accomplishment is important for one's self-esteem and self-efficacy, which in turn have been shown to be related to well-being and effective goal pursuit (Bandura, 1997; Anderson and Honneth, 2005). So, the concern is that improvements in self-regulation achieved through e-coaching could be undermined by the sense that these improvements are not really one's own, but actually the accomplishments of the e-coaching system. In such a scenario, having one's self-regulation supported by an e-coaching system might thus backfire by having a negative effect on one's motivations, which might in turn actually lead to a disadvantage in a competitive environment if one no longer feels confident that one has something to offer.

There is also a deeper concern here—albeit less related to social justice—about becoming alienated from oneself in the sense of no longer feeling that one's actions are a result of one's own motivations. The core value that is at stake here is authenticity: "being one's own person, to be directed by considerations, desires, conditions, and characteristics that are not simply imposed externally upon one" (Christman and Anderson, 2005, p. 3). If e-coaching systems turn out to have the unforeseen effect, at least for a subset of its users, that they will find themselves acting, feeling, responding, thinking, etc. in response to an e-coaching system's suggestions in a way that is incongruent with who they feel they are, then this by itself is cause for sustained further reflection on the value-sensitive design of these systems.

Finally, with regard to feelings of accomplishment there are also

concerns about the deterioration of aspects of our culture as a whole, related to the extent to which other people recognize, acknowledge and give credit for actions that were wholly or in part performed by an agent plus e-coaching system. Consider the parallel case of using a navigation system to navigate towards one's destination. While some people are now beginning to see navigation as a skill that is simply gained with the purchase of a smartphone and data bundle, most people still hold that using a product such as Google Maps is not "real" navigating and attach stigma to being fully dependent on a navigation system. Similarly, then, could being reliant on an e-coaching system for successful self-regulation come to be viewed as a mark of disgrace, leading people to give users of e-coaching systems less opportunities in a competitive environment than those whose self-regulation efforts are unsupported by technology. As such, the concern is that those who need and would benefit the most from e-coaching might get a poor reputation for using such systems and end up being disadvantaged in society because of it. In light of this worry, and given that the sense of accomplishment is essentially contestable and historically changeable, it will be prudent to study these effects in real-life settings (i.e., "in the wild") as e-coaching system are being introduced.

This concludes my survey of some key concerns about social justice that arise with the introduction of e-coaching systems. I do not presume that the set of concerns I have discussed is exhaustive, but I do hope that the issues I have identified offer rich grounds for others for identifying further issues, as well as for further reflection on and investigation into the concerns I have touched upon. For now, let us focus on a second category of concerns that I want to include in the inventory, namely those related to risks of infringements of people's rights of personal autonomy.

3.2 Concerns about Infringing Rights of Personal Autonomy

The concerns in this second category all relate to personal autonomy and the principle of non-interference: being given the opportunities for developing and executing the capacities for choosing and pursuing plans or paths of life for oneself without interference from others. As

mentioned in the Introduction, the concept of personal autonomy itself is a highly contested one that knows many different meanings and has many different aspects to it (cf. Feinberg (1986)). Though I will become more precise in Chapter 4 about how to understand the closely-related notion of self-governance in the context of this dissertation, I will not engage deeply with the various debates about the precise understanding of personal autonomy. Here, the important point to focus on is that personal autonomy is worthy of respect. What respecting personal autonomy entails is captured well by Beauchamp (2005), who writes the following.

> To respect an autonomous agent is to recognize with due appreciation that person's capacities and perspective, including the right to control his or her affairs, to make certain choices, and to take certain actions based on personal values and beliefs. Such agents are entitled to determine their own destiny, and respect requires noninterference with their actions. Respect involves acknowledging decision-making rights and enabling persons to act, whereas disrespect involves attitudes and actions that ignore, insult, or demean others' rights of autonomy. (Beauchamp, 2005, p. 311–312)

In relation to e-coaching systems, concerns about infringing rights of personal autonomy can be divided into two main categories. The first category concerns the *enabling conditions* for personal autonomy, in particular the idea that e-coaching systems may infringe upon aspects of privacy that are necessary for the development and execution of capacities for autonomous choice (Kupfer, 1987; Roessler, 2008). The second category subsequently relates to *procedural independence*, the idea that individuals themselves, and not system designers or algorithms, should be in charge of the life decisions they make. Here, we will thus be concerned with risks of being coerced or manipulated (Yeung, 2017).

Once I have surveyed these concerns, we will stay with the subject of personal autonomy in Section 3.3 but focus our attention on a concern that is not about infringements so much as it is about how the specific interplay between e-coaching technologies and their users may lead individuals to abdicate or neglect responsibilities that they have towards themselves in light of valuing personal autonomy, thereby

failing to meet certain *authenticity conditions* for personal autonomy.[1]

Since the ideas relevant to the third category will be at the heart of my argument for the complacency concern that I will develop in the remainder of this dissertation, they will be discussed as a lead-in to my main argument. First, however, let us turn to the subject of privacy.

Concerns about Privacy

The Charter of Fundamental Rights of the European Union (2000/C 364/01) states that "[e]veryone has the right to the protection of personal data concerning him or her" and that "[s]uch data must be processed fairly for specified purposes and on the basis of the consent of the person [...]" (article 8). As mentioned, the continuous, 24/7 monitoring necessary for effective e-coaching generates massive amounts of data to be collected, processed and stored. The collected data can be sensitive, personal data (especially in health-related or finance-related domains), and this then raises questions about how long the data should be stored, where the data should be stored, how the data should be encrypted, and who should be given access to the data. For some kinds of sensitive, personal data, it may even be questioned whether the data should be collected at all. These aspects of privacy concerns have applicability pretty much across the board when it comes to computing systems that collect data.

Not surprisingly, then, privacy has been discussed extensively in ethical debates about pervasive and ambient computing systems (e.g., Bohn, Coroamă, Langheinrich, Mattern, and Rohs, 2005; Brown and Adams, 2007; Berdichevsky and Neuenschwander, 1999; Karppinen and Oinas-Kukkonen, 2013; Roessler and Mokrosinska, 2013; Yeung, 2017) and is by far the number one concern that is currently making headlines in public debates on these topics. As persuasive systems fall under this general heading, by extension, these privacy issues about the safe collection and storage of personal data also have bearing on persuasive systems and therefore e-coaching systems as well. The specific worries that arises with e-coaching systems in particular, however, have not received much attention in the literature. They will be our focus here.

[1]Notice that the notion of "authenticity conditions" in the context of personal autonomy should be distinguished from the more diffuse sense of authenticity understood as a form of integrity (cf. Christman (2009, p. 134)).

First, with e-coaching systems, there are additional concerns about protecting the conclusions that e-coaching systems draw from connecting and drawing inferences from various data streams. Though it may be argued that many conclusions are already implicit in the data on which the e-coaching system bases itself (e.g., the mean walking distance per day is implicit in a stream of data points about walking over time), the relationships about which e-coaching systems have knowledge can generate additional information that needs to be protected. For example, the mean walking distance per day may well be a neutral measure, but by taking into account a person's goals, an e-coaching system could yield conclusions such as "this man never achieves the walking targets he sets for himself," which can easily be interpreted as revealing a "lack of conscientiousness" or "lack of persistence". For this reason, it is important that the e-coaching systems are designed to deal carefully with both the data they process and the interpretations they suggest, particularly since more conclusions might be able to be derived in the future than is currently possible.

A second privacy-related worry that arises with e-coaching systems has to do with data anonymization. Typically, when data is collected, the strategy to mitigate privacy concerns is to remove (or obfuscate) from the data any "personally identifiable information" such as a person's name, social security number, date and place of birth, etc., but also "linkable" information such as medical or employment records. Unfortunately, research studies have shown that it is often possible to "re-identify" anonymized data to a high degree (e.g., Narayanan and Shmatikov, 2008; De Montjoye, Hidalgo, Verleysen, and Blondel, 2013), especially when there is access to different types of data. As we have seen, e-coaching systems are likely to collect data from a wide variety of sources, which means that the data profile that the system creates will very likely be so specific to the individual that he or she may still be identified, even when the data is anonymized.

In some data collection domains, it is possible to strengthen the anonymization process by only storing data in aggregated form, such that the dataset only contains information about a population, and not about any individual. This kind of privacy protection is suitable for example for municipalities who want to track water usage or electricity: by storing the aggregated usage per area instead of per household, the privacy of each household is ensured. With e-coaching systems,

however, a privacy protection strategy like this would severely limit the e-coaching system's potential. For an e-coaching system needs data on the individual level in order to tailor its advice to the user. For example, to warn a user for potential fast food temptations, or to suggest a good hiking path in the vicinity, it would need to have access to accurate location data, which is notoriously sensitive information.

This tension between effectiveness and the protection of privacy, which seems to be inherent with technologies such as e-coaching systems, calls for sustained critical reflection in light of the continuously changing perspectives on the role that privacy plays—for example in fostering a protective sphere for developing and maintaining personal autonomy, as well as for engaging in social interactions that are integral to a well-functioning society (Regan, 2000; Roessler, 2008; Roessler and Mokrosinska, 2013)—as well as changing regulations. One highly relevant development in this respect has been the ratification within the European Union of the General Data Protection Regulation (GDPR), which formalizes, among other things, the "right to be forgotten" (*right to erasure*, article 17 GDPR). Under this regulation, individuals should at least be given the option to rescind their consent to their data being used, and, at least in theory, to have their data be deleted.[2]

Regardless of how this development further unfolds, however, it is important to note that regulations such as the GDPR should not be seen as a resolution of the above-mentioned concerns. After all, even with regulations in place about data ownership and erasure, it not evident that such regulations in practice will suffice to adequately protect people's privacy (for example, the risk of large-scale data leaks remains very real). Moreover, and perhaps more importantly, general questions remain about the societal desirability of having citizens be tracked on an individual basis, even with their consent, by a wide-ranging group of interested parties, some of whom may have malicious intent. It is these worries we turn to now, about how the collected data, even if it is collected and stored in a secure way, may be misused in attempts at coercion or manipulation.

[2]A recent judgment by the European Court of Justice has significantly weakened the force of this regulation, though, ruling that the right to be forgotten is not an absolute right and that de-referencing of information need only happen to the extent that it can no longer be accessed straightforwardly from within EU member states (European Court of Justice (Grand Chamber), 2019).

Concerns about Coercion and Manipulation

As with other persuasive systems, there are risks of coercion and manipulation associated with e-coaching systems. Though some authors, most notably Fogg and Oinas-Kukkonen, have attempted to explicitly exclude coercive and manipulative systems in their definitions of persuasive systems by adding a clause about voluntary participation of the user (e.g., Fogg, 2003; Oinas-Kukkonen, 2010), Smids (2012) has convincingly argued that simply adding a "voluntariness" clause to a definition does not take away the concerns that arise with these types of systems. It is these concerns I will survey now, beginning with concerns about coercion.

Coercion is traditionally characterized in terms of involuntary action in response to force (e.g., Lucas, 1966; Kelsen, 1967; Lamond, 2000) or credible threat (e.g., Nozick, 1969; Wertheimer, 1987). In relation to e-coaching systems, force is best understood not as physical force—i.e., in the way that being held down is coercion through force—but instead more broadly in terms of having one's freedom restricted. Examples could include systems for stimulating exercise that would deny opening the door to one's smart home unless one had achieved one's activity targets for the day, or systems for improving eating habits that would interact with one's digital wallet to deny payments at fast-food restaurants. Whether these examples truly demonstrate coercion depends of course on a number of factors, including whether the individuals in these cases have given consent to having such measures be imposed upon them as forms of "commitment mechanisms," but the examples should be illustrative nonetheless of the kind of force e-coaching systems could potentially level against their users.

Besides force, coercion can also occur through credible threat (cf. Nozick (1969)). In one of the previous sections, we already encountered a possible scenario in which this form of coercion might occur, namely the scenario in which people would feel pressured to employ e-coaching systems for financial reasons such as obtaining insurance premium discounts. In such a scenario, people might subsequently also feel pressure to follow an e-coaching system's suggestions if their insurance company would threaten to withdraw their insurance premium discount again on the basis of non-compliance. Another example would be if a "freeware" e-coaching system would, after having collected data for some time, threaten to sell or expose

those data unless the user makes a donation to the company behind the e-coaching system.

Equally troubling but potentially more pervasive are the risks of manipulation, since manipulation comes in many forms (Sunstein, 2015). Typical examples of manipulation often involve a form of deception where the manipulator uses false information to motivate his or her target to action. An example of this in relation to e-coaching would be an e-coaching system for sleep improvement that, in order to boost product sales, promotes its own brand of non-evidenced-based "herbal sleep medication" by making wild, unsupported claims about the product's effectiveness. Another example would be a system that would present its user's data in misleading ways in the hopes of motivating its user (e.g., showing a user her data in comparison to fake data from fictive people to show that she needs to do better).

However, as several authors on the subject of manipulation have pointed out, deception is not a constitutive element of manipulation (e.g., Mele, 1995; Noggle, 1996; Cave, 2007; Sunstein, 2015). Mele, for example, articulates manipulation in terms of agents coming "to possess pro-attitudes in ways that bypass their [...] capacities for control over their mental lives" (Mele, 1995, p. 166). Similarly, Cave introduces the notion of "motive manipulation," where the manipulator mobilizes "non-concern" motives of another such as reflexes or biases "so as to induce her to behave or move differently than she would otherwise have behaved or moved, given her circumstances and her initial ranking of concerns" (Cave, 2007, p. 132). Thus, central to these kinds of conceptions of manipulation is not the veracity of the information that is used to induce certain behavior, but the manner in which the behavior is induced. In other words, these conceptions emphasize that it is possible for manipulators to engage in manipulation without deception if they somehow bypass the target person's mental capacities. The hard question, of course, is what constitutes such bypassing.

On the extreme end of the spectrum, there typically is consensus that techniques such as subliminal messaging, brainwashing, or the hypothetical practice of thought injection count as forms of strict bypassing and are therefore problematic in relation to personal autonomy. Moreover, it has been argued that covert, Big-Data-driven "hypernudges" of which individuals have no knowledge also bypass people's mental capacities and should therefore be critically examined and

their use subjected to public debate (Yeung, 2017; Susser, Roessler, and Nissenbaum, 2018, 2019). However, once we move away from the more obvious cases, there is an entire grey area—including accepted social practices such as advertising and sales—in which it is difficult to delineate between manipulative and non-manipulative influences since these influences do engage people's rational capacities, at least to a certain extent.

As foreshadowed in the previous chapter, it is in this category that e-coaching systems fall. They engage with the user through an ongoing, constructive dialogue, while at the same time using persuasive techniques to help with goal striving that the user may or may not pick up on. For instance, consider again the opening example from the Introduction where an e-coaching system, after prior consultation with its user, subtly dims the lights in order to help the user to go to bed. Interventions such as these may escape the user's focal attention in the moment, but then again, could also easily be observed and resisted. Or consider an e-coaching system suggesting to do a workout similar to a workout that the user's favorite movie star (supposedly) does. In this scenario, the system attempts to persuade the user to exercise by playing into the user's bias to care for celebrity endorsements. Again, though, this attempt can be observed, reflected upon, and, under normal circumstances, resisted.[3]

So, while the dialogue between e-coaching system and its user is no automatic guarantee against all forms of manipulation—recall how parties may have incentives to steer the coaching conversation towards the use of certain brand products or towards certain behavioral goals—it does mean that the e-coaching system's influences are generally in the realm of what Kane calls "overt nonconstraining control" (Kane, 1998) against which individuals can, in principle, guard themselves.[4] This presupposes a responsibility of the agent—a responsibility that, as I will argue in Chapter 4, flows from valuing self-governance—to remain vigilant with regard to an e-coaching system's suggestions by critically screening them in light of their own goals, plans and overall values. But what if there is something about e-coaching technologies

[3]On the (ir)resistability of the motives that are appealed to by a manipulator, see Fischer (2004).

[4]I use "generally" here to indicate that, strictly speaking, purported e-coaching systems could implement non-persuasive techniques such as subliminal messaging, though doing so would technically put such systems outside the class of e-coaching systems.

that makes people fall short in exactly this regard? That these technologies facilitate a form of complacency in their users, causing them to be less vigilant than they ought to be? It is this idea that motivates the next section, and that will be the focal point of the remaining chapters.

3.3 Concerns about Unintended Diminishment of Personal Autonomy

As mentioned at the beginning of the chapter, there is a another category of concerns about e-coaching systems potentially diminishing personal autonomy. This category is related to but distinct from the risks of manipulation and arises with the interplay between e-coaching technologies and their users. The central idea is that e-coaching technologies might (unintentionally) facilitate a kind of complacency in its users, so that users fall short in meeting a responsibility that they have towards themselves to keep up a certain level of vigilance with regard to the system's suggestions. These concerns are related to the risks of manipulation in the sense that complacency in relation to assessing outside influences can make individuals more vulnerable to attempts at manipulation. However, they are distinct concerns in that complacency may arise when there is no malicious intent (cf. Noggle (1996)) and no bypassing of the user's mental capacities. In other words, complacency can arise in the absence of any suspicion of manipulation.

In this final section of this chapter my aim is provide a first sketch of the complacency concern and the associated risk to personal autonomy. Before beginning properly, however, let us first briefly consider the word "complacency". In Section 4.2.2 I will introduce a precise definition of the term, but for the time being, I will use it colloquially to denote a state of undeserved or unwarranted self-satisfaction that leads to a lack of action or effort. Thus understood, one might be tempted, in the context of this dissertation, and especially with regard to the notion of self-regulation support systems, to think of complacency foremost as one of the problems that e-coaching systems (or human coaches for that matter) attempt to combat. For example, a person's unwarranted self-satisfaction with his diet might be what is keeping him from losing the weight he wants to lose. If this is indeed

so, why think that complacency is a concern that arises with e-coaching systems?

The answer is that the specific complacency worry that I will be developing is about a kind of complacency *in relation to certain aspects of our practical reasoning* that is facilitated by e-coaching systems. That is, I am concerned with a kind of complacency that may arise with the use of e-coaching systems, even while sometimes, in parallel, the e-coaching system might be helping to overcome a user's complacency in relation to a particular behavioral goal. With that clarificatory note in place, let me begin laying out why I believe complacency in relation to practical reasoning may occur, and why it would be a problem for personal autonomy.

The first observation is that people have various cognitive biases that, normally, allow them to be cognitively prudent in their reasoning, but that also systematically lead them to make certain kinds of mistakes in their decision making. Well-known examples include the "status quo bias" where people disproportionately stick with the status quo (e.g., Samuelson and Zeckhauser, 1988), the "risk aversion bias" where people prefer avoiding losses as opposed to acquiring equivalent gains (e.g., Tversky and Kahneman, 1991), and "confirmation bias" where people seek or interpret evidence "in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" (Nickerson, 1998, p. 175). While there is discussion about the utility or disutility of these and other biases, the empirical evidence for their existence abounds.

In much the same vein, it has been shown that people also have a bias in their interactions with automated systems. In Section 4.2.1 I will review some of the key literature describing this "automation bias," but here it suffices to report the upshot, namely that people have a tendency to "fail to notice problems because the automation does not alert them, [...] [or] erroneously follow automated directives or recommendations" (Cummings, 2004, p. 2). As we will see, this issue is related to a systematic overestimation of the accuracy and reliability of automated systems. This presumption of accuracy and reliability, coupled with the widely-held beliefs that computers enjoy a sense of objectivity and are computationally superior to human beings, can lead to over-reliance on automated systems, which in turn can draw out certain kinds of errors. Such errors may include but are not limited to failing to notice shifts in one's values or priorities or

following suggestions that are not aligned with one's values and may invoke a sense of alienation.

Certainly, not all errors resulting from over-reliance on automation stem from instances of complacency. Sometimes, choosing to rely excessively on a system that is known to be reasonably reliable but fallible is the rational thing to do, especially in scenarios with multiple tasks that require an effective use of cognitive resources. Errors as a result of over-reliance may then still occur, but those should not be attributed to complacency as there is no unwarranted self-satisfaction involved.

My contention, however, is that users of e-coaching technologies will also be making errors following interactions with their e-coaching systems that can be traced to a lack of effort resulting from an unwarranted self-satisfaction with their accomplishments in relation to processes that are essential to personal autonomy, such as ensuring that one's values, goals, beliefs, intentions and plans fit together in a coherent and consistent way, and that their actions are "linked up" in the right way to this constellation of mental states. Put differently, the worry is that people's interactions with e-coaching systems will lead users to culpably overestimate their accomplishments with regard to processes essential to personal autonomy (the state of unwarranted self-satisfaction) that in turn will lead them to culpably drop their vigilance with regard to those very processes. Clearly, there is more to be said about this, and I will elaborate on these points in the chapters to come. What I want to draw attention to for the present discussion, however, is the thought that dropping one's vigilance in the way just described is indeed problematic for personal autonomy, regardless of the specific impact of each complacency-related error.

As mentioned before, an important premise here is that personal autonomy does indeed demand some level of vigilance with regard to certain processes. The processes I have in mind are those that pertain to the agent determining where he stands, viz. the agent's practical standpoint (cf. Dworkin (1976); Frankfurt (1987); Dworkin (2015)). I will say more about this in Section 4.3.1, but in short, the thinking is that if personal autonomy requires making up one's own mind about how to live one's life, then surely part of making up one's mind is performing "due diligence" with regard to scrutinizing both internal influences (e.g., a desire to ϕ) and external influences (e.g., peer pressure to ϕ).

Though different accounts of personal autonomy may diverge with regard to the standards of vigilance that agents should to held to, I take this premise, in its most minimal form, to be broadly compatible with various accounts of personal autonomy.

To illustrate the kind of vigilance I am concerned with, consider again advertisements and other marketing strategies aimed at convincing people to buy certain products. Using catchy slogans and attractive commercials, many marketing strategies attempt to exploit biases that consumers are known to have. These attempts by themselves do not undermine people's personal autonomy per se, because agents can guard themselves against these exploits by critically scrutinizing their perceptions (e.g., an attractively packaged product) and their response to what they perceive (e.g., a desire for owning that particular product). Agents who guard themselves in this way are certainly not immune to influences—for example, certain commercials may appeal to stereotypes that agents unconsciously endorse—but armed with self-knowledge and a dose of awareness, these agents are at least in a position to counteract those influences by giving less weight to the reasons they provide.

On the extreme opposite end of the spectrum, it would seem that agents who let their shopping behavior be determined fully by advertisements and commercials without further reflection are not acting autonomously. We may liken these kinds of agents to Frankfurt's "wantons," who let their actions be guided solely by first-order desires (Frankfurt, 1971). Of course, in actuality, the extent to which people are susceptible to commercials and have the capacity to successfully guard themselves against these influences will be a matter of degree, but that is not the point. Rather, the point is that in order to act autonomously, there is some threshold level of vigilance required. If this point is accepted, then we can see exactly why the complacency concern should be carefully considered in the debate about e-coaching systems. For if there is a risk that e-coaching systems facilitate a complacency in their users with regard to processes that are essential to personal autonomy, then that effectively threatens their personal autonomy.

This, in a nutshell, is the complacency concern. Now that we have a first articulation of it, we can see more clearly that what is characteristic about this concern is that it is less about personal autonomy not being respected per se (cf. the concerns from Section 3.2), and

more about people being lulled into a state in which they lose or at least reduce their personal autonomy by neglecting to engage appropriately in processes that are essential to acting autonomously. Note, however, that though the emphasis on agential responsibilities in this category of concerns marks a departure from other concerns I have surveyed in this chapter, it should not suggest that the full responsibility for counteracting automation-related complacency lies with the individual user. As I will argue in Chapter 5, designers and developers of e-coaching systems carry a responsibility in this regard as well. However, here (and in the next chapter), I want to emphasize that individuals have a role to play in the responsible use of e-coaching systems. This is a perspective that has remained underexposed in much of the literature on the ethics of this class of self-regulation support systems.

In the remainder of this dissertation, my aim will be to further unpack the complacency concern and to shed light on its implications. My approach to making headway with this project is to zoom in on one particular model of autonomous agency in order to have a more specific framework in which to frame and assess the complacency concern. I will introduce this model of autonomous agency at the beginning of Chapter 4. Then, in the sections thereafter, I will use that framework to revisit and expand on the thoughts from the present section about both the risk of complacency being facilitated by e-coaching systems, and the consequences that this may have for people's personal autonomy. Finally, in Chapter 5, I will provide suggestions for how to manage the risks.

Chapter 4

Complacency, Diachronic Self-Governance, and E-Coaching

"[I]n diachronic self-governance one is, as it were, acting 'together' with oneself over time."

Michael E. Bratman

The emergence of e-coaching systems brings with it a risk to people's personal autonomy because such systems facilitate complacency with regard to processes that are essential to governing one's own life. That, at least, is the concern I sketched at the end of the previous chapter. In the present chapter, my aim is to work out the complacency concern further by explicating the concept of complacency and explaining how complacency in relation to practical reasoning undermines personal autonomy. To do so, I will need a particular model of self-governing agency and corresponding vocabulary in order to frame the concern and analyze its implications more precisely.

Since the complacency concern itself is not specific to any one model of agency, it could, in principle, be worked out in different ways using different models. This does not mean, however, that the choice for a particular model has to be arbitrary. In choosing my model, I am guided by four desiderata. The first desideratum is that the model of self-governing agency should be compatible with the kind

of theorizing found in psychology—specifically with self-regulation theory (see Chapter 1)—where the main postulates are processes, states and events whose specific interplay make up the agent. For such continuity in theorizing will allow me to say more about how self-regulation relates to self-governance (see Section 4.1).

The second desideratum is that the model should be compatible with a degree of influence from and reliance on the outside world. That is, the model should not rule out on principled grounds the possibility of self-governance for individuals who determine where they stand, in part, by engaging with other individuals or with technologies such as e-coaching systems. This is important for compatibility with recent theorizing about the social aspect of personal autonomy, emphasizing that autonomous beings need advice from others, and shape their thoughts, characters and capacities in social communities (e.g., Christman, 2004; Westlund, 2009; Mackenzie, 2008; Oshana, 1998). And this should ring true, even for those who do not agree that "autonomy is itself a socially constituted capacity" (Mackenzie, 2008, p. 519). In much the same vein, the literature on enactivism and embedded agency more generally (e.g., Varela, Thompson, and Rosch, 1991; O'Regan and Noë, 2001; Noë, 2004; Clark, 2011; Hutto and Myin, 2013; Gallagher and Bower, 2014) give force to the thought that interacting with structures in the environment is part of the very fabric of our being. Taken together, these ideas suggest that an account of self-governance should leave open the possibility that users of e-coaching systems can be self-governing.

The third desideratum is that the model should capture the temporal nature of self-governing agency. This desideratum is motivated by the temporally extended nature of self-regulation support by e-coaching systems. For these systems will be employed with the aim of assisting people in their efforts to coordinate their actions over time in service of striving towards some self-chosen end (e.g., trying to quit smoking, or getting sufficient sleep). This means that users of e-coaching systems are likely to outsource certain processes to these systems on a continuous basis for sustained periods of time. As such, we will want to understand not only how complacency affects whether a particular action is self-governed or not (*synchronic self-governance*) but also the way in which complacency affects self-governance over time (*diachronic self-governance*).

Finally, and related to the previous point, the fourth desideratum is that the model offers an explicit account of the roles that intentions, plans and planning play in self-governing agency. This is pertinent to the discussion about e-coaching systems because, as discussed in Chapter 2, these systems will support their users in their efforts to coordinate their actions over time by engaging with them in a collaborative conversation to provide assistance with deliberative planning processes, for example by proposing options for action, helping to schedule one's activities, or making suggestions with which to form implementation intentions. Given that we want to understand the implications of overlying relying on this kind of planning assistance, it is therefore important to be explicit about how the process of planning and the resulting plans feature in how we determine for ourselves how to live our lives.

The model that meets these four desiderata particularly well is Bratman's planning theory of agency (Bratman, 1987, 2014, 2018). As we will see, this theory offers a naturalized, event-causal account of self-governing agency (desideratum 1) that allows for influence from and reliance on the outside world (desideratum 2) and explains the way in which human beings determine and coordinate their actions across time (desideratum 3) by appealing to the roles of intentions, plans and planning (desideratum 4). Since I believe meeting these desiderata makes a model of self-governing agency especially well-positioned for the kind of analyses we are after, I will use the planning theory throughout this chapter and the next. That said, let me reiterate my contention that complacency will likely be a concern in other theories of self-governing agency as well, so that the general points I make about complacency (e.g., in Section 4.2.2) will still be relevant to readers who favor a different account.

With these preliminaries now out of the way, let me outline the structure of the chapter. First, I will introduce the planning theory of agency and unpack a number of key concepts of that theory such as intentions and plans, and the rational pressures towards consistency and means-end coherence. In addition, I will outline the planning theory's view on personal autonomy, which will be expounded in terms of self-governance. Once this framework is in place, I will return to the complacency concern. Specifically, in Section 4.2 I will argue for the plausibility of the empirical claim that e-coaching systems bring

with them a heightened risk of complacency. I will do so by reviewing empirical literature on "automation-related decision bias" and, after defining complacency more precisely, arguing that the findings reported in that literature offer support for the claim. Next, I will argue in Section 4.3 for the conceptual claim that complacency with regard to processes that are essential to practical reasoning indeed affects self-governance. Finally, to conclude the chapter, I will introduce and discuss a series of fictive examples in Section 4.4 to illustrate different ways in which complacency may manifest itself in one's practical reasoning and the effect this has on one's personal autonomy.

4.1 The Planning Theory of Agency

Central to the planning theory of agency is the observation that human beings are not only purposive agents in the sense that they have the capacity to act[1], but that they are also "planning agents" in that they have the distinct capacity to coordinate their actions and activities across time by forming future-directed intentions and plans of action (Bratman, 1987, 2007). As noted in Chapter 1, this kind of temporally extended coordination of action is pervasive in the everyday life of human beings and essential to successful self-regulation.[2]

[1]Acting in this context is standardly construed in terms of intentionality and is therefore to be distinguished from merely displaying behavior. Different accounts have been proposed of the intentionality of action (e.g., Anscombe, 1963; Davidson, 1963; Goldman, 1970). On the standard theory of action, which originated with seminal work from Davidson, the intentionality of action is explained in terms of the action being caused (in the right way) by the agent's mental states (see Davidson, 1963). For example, one's "turning off the light" when going to bed is an action, under this description, insofar as the moving of one's arm to flip the light switch is caused by a pro attitude (e.g., a desire) "to turn off the light" and a belief that flipping the light switch is a means to turning off the light. This pro attitude, together with the belief, form the primary reason for moving one's arm to do the flipping. In contrast, were one to hit the switch accidentally because one moved one's arm in a fright reflex, the flipping of the switch would not be one's action, even though one's behavior causes the light to turn off.

[2]Notice that the focus on intentions and plans of action does not mean that planning theorists diminish or underestimate the role that habits, tendencies, reflexes and emotions play in people's behavior. For example, choosing between two equally good options (e.g., different types of cereal in the supermarket) will often "not be explainable by [...] desires, beliefs and intentions [but] presumably [...] at some other—perhaps neurophysiological—level" (Bratman, 1987, p. 11). Moreover, it should also not suggest that agents are continuously using their precious cognitive resources for planning. Rather, as we will see, the planning theory neatly helps explain that agents can use their

The planning theory proposes to explain the pervasiveness of planning in human agency by seeing plans and plan-like states such as intentions as integral to an agent's psychic economy. Thus, on the planning theory, planning states are thought to be irreducible *sui generis* psychological states[3] that distinguish themselves from beliefs and desires by being reasonably stable, by entailing commitment, by being action-guiding, by being subject to (internal) norms of consistency and coherence, and by the filtering role they play in an agent's practical reasoning. A few of these characteristics of planning states we already encountered in Section 1.2.1, but I will unpack these notions here more fully, beginning with the stability of plans and intentions.

The stability aspect of planning states relates to the idea that once agents form a plan or intention, they have, in principle, made up their minds about what they will do. Having an intention at time t_0 to ϕ at time t_1 entails a commitment to ϕ-ing at t_1.[4] This intention is of course open to reconsideration under certain conditions (sometimes situations turn out differently than expected) but the guiding idea is that, normally, a future-directed intention will be retained until the time to act arrives.[5] As Bratman puts it, "the agent retains that intention over time by way of nonreflective, reason-preserving non-reconsideration" (Bratman, 1987, p. 92). When the time to act arrives (t_1), it is at that point that the future-directed intention becomes a present-directed intention that causes and guides the action. In this sense, future-directed intentions can be said to be action-guiding.[6]

Plans, which for Bratman are "intentions writ large," (Bratman, 1987, p. 29) share a similar sort of stability over time, in that, while they are subject to revision, they too characteristically resist reconsideration. Normally, that is, a planning agent will follow through on a plan once the plan has been formed.[7] With plans, however, following through

cognitive resources effectively by settling in advance on a course of action, so that that they can follow their predetermined plan of action semi-automically when the time to act arrives.

[3]For a contrasting view in which intentions are thought to be reducible to beliefs and desires, see for example Sinhababu (2013).

[4]In this respect, intentions are very different from desires, as desires may be fleeting and non-committal to any action.

[5]For discussion about the rationality of such reconsiderations, see for example Bratman (1987, p. 68), and also Holton (2009, p. 75).

[6]For more on the notion of action guidance, see Frankfurt (1978).

[7]Notice that the characteristic stability of plans leaves open the possibility for plan-

is often not just a matter of acting when the time comes to act, but frequently also requires further deliberation to settle on the means for reaching one's end. For plans are typically partial and hierarchical: they usually start out in a fairly abstract form that agents will make more concrete by forming embedded sub-plans about means as time goes by and they see fit.

Consider for example an agent who has formed the plan to visit Ottawa next fall. This plan is not very specific yet as to travel dates, modes of transportation, or accommodations in Ottawa, but it does entail a commitment to go. As such, having this plan puts rational pressure on the agent to fill in the details of how to realize the visit. Filling in these details need not be done right away; the agent may choose to wait until a later time when he is in a better position to make certain decisions. However, unless the agent abandons the plan, the relevant decisions will have to be made at some point, at least before, by the agent's own lights, it is too late (e.g., when the agent judges that the airline fairs are no longer affordable for him).

This rational pressure on agents to fill in the means to an adopted end is known as the pressure of *means-end coherence*, and this is articulated by Bratman in the statement that it is pro tanto irrational for an agent to be "intending E while believing that a necessary means to E is M and that M requires that one now intend M, and yet not now intending M" (Bratman, 2018, p. 78). So, if E is "visiting Ottawa" and M is "buying an airplane ticket to Ottawa," then for an agent who intends to visit Ottawa and who believes that buying the airplane ticket now is necessary for visiting Ottawa, it is pro tanto irrational not to now intend to buy the airplane ticket.

In addition to means-end coherence, the planning theory also holds that planning agents are subject to a rational pressure towards *consistency* of all of one's planning states.[8] That is, an agent's plans and intentions should not be incompatible with one another (intending to ϕ and intending to ψ, while believing that ϕ and ψ are not co-possible).

ning agents to change their minds: if agents do reconsider an existing plan, taking into account the information that they have now (as opposed to the information they had earlier), it is entirely possible for them to land on a different outcome. The point about stability is rather that agents will typically not reopen matters that have already been settled.

[8] For a discussion about grounding the pressure for consistency, see Bratman (2018, pp. 38–42).

For example, being in Ottawa in the fall is incompatible with teaching a course in the Netherlands during that same period, and it would therefore be pro tanto irrational for an agent to plan to visit Ottawa and to plan to teach a course in the Netherlands in the fall period, while believing that it is not possible to realize both.

The picture that emerges is one where prior plans and intentions, as long as they are not reconsidered or abandoned, help structure an agent's practical reasoning "in ways that are shaped by requirements for consistency and coherence of those intentions and plans" (Bratman, 2007, p. 290). In other words, when an agent is deliberating about a new intention or action, his prior plans and intentions form a temporarily fixed background that constrains deliberation and serves as a "filter of admissibility" (Bratman, 1987, p. 33) for the options that the agent can consider without being either inconsistent or means-end incoherent. So, the agent who plans to visit Ottawa in the fall has his deliberation constrained by that very plan such that he considers it rational for him to intend some means that will enable him to get to Ottawa (e.g., buying an airplane ticket) and to ignore options that are not co-possible with the background plans such as teaching a course in the Netherlands at the same time.[9]

Together, these rational pressures towards coherence and consistency of one's plans against the background of one's beliefs help explain why planning is, as we have seen in Chapter 1, an important part of self-regulation. For once an agent adopts a goal intention (in contrast to merely desiring a goal), the agent is under pressure to further deliberate about coordinating this new planning state with prior and future plans (what Bratman calls "plan agglomeration") and to consider (and intend) the means to achieving the goal. In this way, planning helps agents with their effective goal pursuit.

Naturally, though, the further question arises what the sources of these rational demands are. In the literature, authors such as Harman, Velleman, and Wallace have suggested that these demands on planning states are actually derived from norms about beliefs (Har-

[9]The reasoning need not stop here, of course. The agent may well plan to teach a course in the spring, since that is perfectly co-possible with his visit to Ottawa in the fall. In addition, the agent may also rationally reconsider his plan to visit Ottawa in light of his teaching obligations. This is besides the point of the example, however, since the example is meant to show precisely that prior plans, as long as they are not reconsidered or abandoned, constrain deliberation.

man, 1976, 1986; Velleman, 2000; Wallace, 2001). While their specific accounts differ, the basic idea is that inconsistency or incoherence among one's intentions always involve inconsistency or incoherence among one's beliefs, and that the norms of consistency and coherence for beliefs are more fundamental than those for intention. Bratman, however, rejects this idea and instead argues that it is "the complex of *practical* roles of intention that lies behind these norms" (Bratman, 2018, p. 123, his italics) and that planning agents have a reason of self-governance to be responsive to these norms. Such responsiveness "will tend to support the effective pursuit of one's ends" (Bratman, 2018, p. 123) as well as essential forms of sociality without which human life "would be impoverished and difficult to understand" (Bratman, 2018, p. 112).[10] Therefore, Bratman argues, "our reason for conforming to these norms of practical rationality derives in part from our reason to govern our own lives" (Bratman, 2018, p. 77). To better understand this argument, it is important to first understand how self-governance is characterized on the planning theory, beginning with what it means to be self-governing at a time (synchronic self-governance).

4.1.1 Synchronic Self-Governance

The guiding idea for the planning theory's characterization of self-governance at a time is that it involves guidance of one's thought and actions by one's relevant practical standpoint. Since the planning theory supposes that agency is made up of structures of and the interplay between mental states (beliefs, desires, intentions, plans, etc.) and associated reasoning processes—in short, since it is a naturalized, event-causal theory of agency[11]—this idea is to be expressed in terms

[10] The thought here is that an agent's capacity to form coherent and consistent plans also supports shared activity through coordination of actions between agents. If one wants to move apartments, for example, one will have to interpersonally coordinate with landlords, friends, family, professional movers, etc. to ensure access to the new apartment and that the apartment is painted before furniture is put in place. This kind of shared activity is not the focus of the present discussion, so I will put it to the side here, but see Bratman (2014) for a planning account of shared activity.

[11] Though the event-causal framework is the most widely accepted view in the philosophy of action, objections have been voiced by agent-causation theorists who hold that the reductive nature of the event-causal framework fails to explain how an action is really an agent's intervention, and not just another type of event that happens to occur. Their alternative proposal involves agents as substances that cause events. I cannot do justice to this discourse here, but for discussion, see for example Velleman (1992); Mele

of psychological structures that are such that when they guide thought and action, the agent can be said to govern. After all, there is no further, irreducible agent who is pulling the strings and whose judgement is authoritative.

To get to such psychological structures that have agential authority, the planning theory draws on work from Frankfurt who has famously made the case that structures of mental states (on his view, a hierarchy of desires) can represent where an agent stands (Frankfurt, 1971, 1982). As an illustration to Frankfurt's view, suppose an agent has both a desire to go to bed because she is tired, and a desire to stay up late because she is reluctant to go to a difficult meeting that is scheduled for the following morning. If the agent is to do one or the other out of free will, she will have to decide between these two desires and *endorse* one of the two. According to Frankfurt, this is done by forming a second-order volition that determines where the agent stands. In casu, the agent may find it important to make a good, well-rested impression at the meeting, so she desires to have her desire to go to bed on time motivate her action.

In a similar fashion, the planning theory also utilizes a hierarchical structure by supposing that agents can form higher-order planning states that are not about actions as such, but that specify the sort of reasons to give weight to in their practical deliberation.[12] In Bratman's words, these higher-order, *self-governing policies* "concern the significance that is to be given to certain considerations in our motivationally effective practical reasoning concerning our own conduct" (Bratman, 2007, p. 167). Such higher-order policies constitute, at least in part, where the agents stands in relation to certain motivations.[13] For example, an agent might have a self-governing policy "to get sufficient sleep" that gives weight to all considerations having to do with get-

(2003, ch. 10); Lowe (2008, p. 159–161).

[12]Bratman argues that there are pressures from two distinct directions towards such a motivational hierarchy. The first concerns the idea that autonomous agents not only govern their actions, but also the practical deliberations preceding their actions; the second has to do with agents understanding that their practical reasoning is governed by what they care about. For more on this subject, see Bratman (2007, p. 164 and beyond), and also Bratman (2007, pp. 184-185).

[13]In a way, this connects to arguments made by psychologists such as Brian Little that our identities and personalities should not be equated with our genetic predispositions and trait-like tendencies, but that they are co-constituted by our personal projects (e.g., Little, 2014).

ting sufficient sleep. In considering to watch yet another episode of her favorite television series, it is this self-governing policy that puts weight on the alternative option of going to bed. The self-governing policy thus represents a kind of valuing by the agent (Bratman, 2007, pp. 64-66).[14] If, moreover, such valuings play their proper roles in the cross-temporal organization of the agent by way of creating and maintaining connections and continuities of the agent's personal identity over time, they have a presumption to "speak for the agent", viz. have agential authority, such that when they guide action, the agent directs.[15]

It is here that I want to briefly revisit my remarks from Section 1.2.1 where I contended that certain kinds of (plan-related) reflections are outside of the scope of self-regulation. The rationale for this line of thinking is that we want to avoid ending up with a bloated, overly inclusive conceptualization of self-regulation such that all forms of practical deliberation fall under its heading. Now that we have the basic constructs of Bratman's conceptual framework in place, I think we are in a position to see at least the outline of my proposal, namely that self-reflection and planning are only part of self-regulation insofar as these processes are directed at deliberating instrumentally about the means necessary for meeting certain ends. Presumably, this may sometimes involve deliberating about and setting intermediate goals in pursuit of an overarching end, but it does not extend to deliberating about what Richardson has called our "final ends" (Richardson, 1997). As such, we should consider higher-order reflections—including those about one's life values, the non-instrumental ends that one finds worthwhile pursuing, and the kind of considerations one plans to give weight to in one's practical reasoning—as endeavors distinct from self-regulation as they have to do with determining one's practical standpoint rather than with striving towards any particular end.

Getting back to the main line of discussion, we should note that having self-governing policies in place is not by itself sufficient for self-governance. For an agent to be self-governing according to the planning theory, three more conditions must be met. First, the higher-

[14] As Bratman is quick to point out, valuing of this kind is to be distinguished from judging what is good. See Bratman (2007, p. 172 fn. 26).

[15] For elaboration, see for example Bratman (2007, p. 207–208). Notice that the agential authority of our valuings does not originate from their place in the hierarchical structure but from the role that they play in our identities.

order policies that concern practical reasoning have to, at least in part, also concern their own role in the agent's practical reasoning. This also relates to the planning theory being a naturalized, event-causal theory, as there is no further agent to endorse the higher-order policy. Rather, the endorsement is built into the policy by it also being an expression of the agent's valuing that her practical reasoning is guided by that very policy. In this sense, some self-governing policies are reflexive.[16]

Second, self-governance requires not just having reflexive, self-governing policies, but also being guided by them in "the appropriate way". As Bratman writes:

> [s]elf-governed agency is agency that is deliberatively guided in an appropriate way by attitudes that, in that context, speak for the agent and have agential authority. In that context, the guidance by those attitudes constitutes the agent's governance of action. (Bratman, 2007, p. 129).

It is important to recognize that Bratman distinguishes here between agential direction and agential governance. We can speak of the former when "there is sufficient unity and organization of the motives of action for their functioning to constitute direction by the agent," but only of the latter when "agential direction [...] appropriately involves the agent's treatment of certain considerations as justifying reasons for action" (Bratman, 2007, p. 177). What this comes down to is that the self-governing policies should structure the framework of an agent's practical reasoning and should play their roles in the agent's weighing of reasons. Put differently, for an agent to be self-governing when she acts, her action should not just be in line with her self-governing policies, but also because of those policies.

Third, and finally, self-governance requires a certain stability in an agent's self-governing policies. Of course, it also expected of self-governing agents that they reflect on a regular basis on their plans, and this includes self-governing policies as these are not immune to rational revision (Bratman, 2007, p. 36). By itself, revising one's self-governing policies is a natural aspect of planning agency, and one that is generally unproblematic for self-governance over time (diachronic self-governance). However, for synchronic self-governance, it is important that there are no conflicts between self-governing policies, i.e.,

[16] See also Harman (1999).

that an agent does not have self-governing policies to both favor considerations about X and considerations about not X. This is because such conflicts would undermine where the agent stands, and would therefore undermine the claim to agential authority.

So, what is key is that agents develop self-governing policies that do not directly challenge each other. On the planning theory, this criterion is fleshed out in terms of the Frankfurtian idea of *satisfaction*, which is "a state of the entire psychic system—a state constituted just by the absence of any tendency or inclination to alter its condition" (Frankfurt, 1992, p. 104). Such satisfaction of the psychic system is thus required for self-governance to ensure that the agent is guided by self-governing policies that establish where the agent stands, and from which the agent is not estranged (Bratman, 2007, p. 203).

Piecing these elements together, we can thus characterize synchronic self-governance on the planning theory as being guided in one's thought and action by higher-order, reflexive planning states that have agential authority and that one is satisfied with.[17] This characterization aims to capture "garden-variety self-governance," in the sense that most planning agents will govern their actions at least some of the time in everyday life. For instance, it is entirely plausible that the agent who plans to visit Ottawa in the fall is synchronically self-governing when he books his hotel accommodation, provided his action is guided in the right way by relevant self-governing policies (e.g., to give weight to considerations about experiencing different cultures, to give weight to considerations about taking time off from work, and to give weight to considerations about being financially responsible) and that there is no tendency or inclination to change those policies.

Vice versa, we can see that synchronic self-governance would be blocked when the claim to agential authority by the relevant structure of psychic elements is undermined. When is this the case? Well, when one's relevant plan states in the current context are inconsistent or incoherent. Consider:

[17]Notice that this characterization qualifies as a *proceduralist* account of personal autonomy, in that it focuses on procedural requirements for being autonomous, rather than on the substance of what is chosen. For a discussion on the different kinds of accounts that can be given, see Anderson (2013, p. 7–8).

> [W]hat is needed for self-governance is guidance of thought and action by a practical standpoint that constitutes where the agent stands. But if relevant plan states are inconsistent or incoherent, then there will not be a fact of the matter about where the agent stands on the relevant practical issue. (Bratman, 2018, p. 126)

We are thus now in a position to better understand the argument that an agent has a reason of self-governance for conforming to the demands of consistency and means-end coherence. For failures of consistency and coherence undermine the agent's practical standpoint, which in turn blocks synchronic self-governance. Therefore, provided an agent cares about self-governance, the agent has reason to conform to the demands of consistency and coherence.

Let us take stock. We have seen that the planning theory of agency holds that human beings are planning agents who engage in temporally extended projects and forms of social activities by coordinating their actions, both intrapersonally and interpersonally. We have also seen that planning states fulfill a filtering role in one's further practical deliberations by posing practical problems about means to ends as well as about co-possiblity of one's overall plans, and that these problems originate from distinctly practical rational pressures of coherence and consistency that planning states, contrary to desires, are subject to. Further, we have seen that planning agents have a (defeasible) reason to conform to these practical norms in individual cases, namely a reason of self-governance.

Finally, we have seen that, on the planning theory, synchronic self-governance is conceptualized as being guided in one's thought and action by higher-order, reflexive planning states that have agential authority and that one is satisfied with. This thus provides an account of what it is for an agent to govern one's thought and actions at a time (though, technically, during small periods of time). What is still missing, however, is an account of what is for an agent to govern one's thought and actions across time. Let us therefore fill this lacuna and bring this section to a close by considering what diachronic self-governance looks like on the planning theory.

4.1.2 Diachronic Self-Governance

For Bratman, there are two guiding thoughts with regard to diachronic self-governance. The first is that it should involve continuity of self-governed choice and action over time as this supports what is "constitutive of the identity of the person over time" (Bratman, 2018, p. 144). Making such continuity constitutive of self-governance blocks a kind of agency that intuitively seems in tension with diachronic self-governance, namely the "lurching from one plan-like commitment to another incompatible commitment seen as equal or incomparable, in a way that involves abandoning one's prior intentions" (Bratman, 2018, p. 143).

On the planning theory, diachronic self-governance is thus spelled out in terms of "synchronically self-governed choices at the relevant times along the way of a relevant planned temporally extended activity, where these choices over time are tied together by the interconnections characteristic of planned temporally extended activity." (Bratman, 2018, p. 227). These interconnections include typical aspects of planning agency such as cross-references between plans at different times (e.g., today's plan implicitly cross-referring to yesterday's plan), interdependence between plans at different times (e.g., the success of a plan for tomorrow depending on the success of today's actions), and the fitting together of sub-plans over time. In a way, Bratman argues, achieving this kind of continuity is analogous to a certain extent to acting together with oneself over time (Bratman, 2018, see pp. 230–236).

The second guiding thought is that diachronic self-governance is also an end that is rational for planning agents to have and to appeal to in their practical reasoning. Consider:

> Rational End of Diachronic Self-Governance (REDSG): If S is a planning agent who is capable of diachronic self-governance then it is *pro tanto* irrational of S to fail to have an end of diachronic self-governance. (Bratman, 2018, p. 220)

Bratman considers this a weak principle, in that "it does not require, even *pro tanto*, that the end of diachronic self-governance be preeminent within the agent's standpoint" (Bratman, 2018, p. 220). This means that different agents might give different relative weight to the end of diachronic self-governance within their standpoints. Still, the think-

ing is that this end does put normative pressure on an agent "in favor of satisfying constitutive conditions of self-governance" (Bratman, 2018, p. 220). As such, this end can, in some instances of temptation—consider the various cases we have seen in Chapter 1—induce pressure towards sticking to one's prior plan by re-shifting one's standpoint at the present moment away from the temptation and back towards one's prior plan (Bratman, 2018, p. 216).

While there are a number of difficult issues that Bratman addresses in his latest work on this subject, I will not engage with those issues here. For the most important take-away for the purposes of this chapter is already evident, namely that self-governance can be diminished not only by synchronic inconsistency and incoherence, but also by incoherence in the interconnections between one's planning states over time. And with that insight, we can start to see better what the issue with complacency is: it is not about the laziness of "going along with a suggestion" every now and again, but rather about being negligent towards oneself by not engaging in the ongoing process of determining one's own synchronic and diachronic standpoint.

Let us therefore shift our attention back to the complacency concern. As already alluded to in Section 3.3, the argument that e-coaching systems facilitate a kind of complacency that undermines self-governance consists, at bottom, of two separate claims. The first claim concerns the ways in which support technologies affect the decision making of the users who employ these technologies. It holds that there is some causal mechanism by which e-coaching systems affect people's reasoning and decision-making processes about what to do. As this claim supposes that certain causal relationships exist in the world, it is, in essence, an empirical hypothesis. As such, it will be appropriate to provide a foundation for the claim by looking at existing empirical literature, which we can find in the neighboring field of decision support systems. Reviewing the literature on *automation bias* and *automation-induced complacency* in Section 4.2 will bring out that the data support the idea that technologies trigger mechanisms by which people make different types of decision-making errors than the type of errors the system was designed to help reduce. The brief review will also bring into focus that there is confusion in the literature about how the term complacency is best understood. In an effort to deal with this confusion, I will draw on work from Jason Kawall on complacency

(Kawall, 2006) to propose a way of interpreting the empirical findings that makes sense conceptually.

The second claim is of a different nature than the first. Rather than putting forward an empirical hypothesis, it makes a conceptual statement about the relationship between practical reasoning and self-governance. This second claim posits that the concepts of practical reasoning and self-governance are dependent, in that self-governance *implies* an appropriate level of vigilance in one's practical reasoning. A claim of this kind needs to be argued for, and this is what I will do in the second part of the chapter. The leading idea there will be that complacency in practical reasoning is a failure to be adequately vigilant with regards to making sure that one's planning states are consistent and coherent, and that one's actions are in line with and because of one's values. Once I have laid out the main argument, I will tease apart the different ways in which complacency in practical reasoning can take shape. Finally, using constructs from the planning model of agency and a series of fictive examples, I will illustrate how these different ways in which complacency may manifest itself in one's practical reasoning can lead to scenarios in which people ultimately regret not having been more vigilant.

If I am right about both claims, then this has normative implications for both the design and use of e-coaching systems. In Chapter 5 we will turn our attention to the normative implications of taking the complacency concern seriously and look at possible mitigation strategies. First, however, let us examine both claims in more detail.

4.2 Heightened Risk of Complacency

The first claim can be formulated as follows.

> **The Heightened Risk of Complacency (HRC) Claim**
> *E-coaching systems heighten the risk of people becoming complacent in their practical reasoning.*

This HRC Claim, as stated at the end of the previous section, is an empirical claim. Ideally, such claims are supported by evidence from randomized controlled trials in which a particular isolated mechanism is manipulated in one group of participants (the experimental group), and not in another group (the control group). At present, no

randomized controlled trial has been done with e-coaching systems that focuses on identifying the type of decision bias that I am claiming exists.[18]

Luckily, automated decision support in itself is not new; decision support systems (DSS) have been around for a number of decades for professionals working in clinical health care (e.g., diagnostic tools) as well as in aviation (e.g., air traffic monitoring systems).[19] My aim for the following subsection, then, is to review evidence from that domain to make plausible the idea that technologies can trigger a particular kind of decision bias.

Of course, doing so requires the further assumption that findings in that related but distinct domain of DSS are likely to carry over to the one we are interested in. I think this assumption can be defended straightforwardly. For one, DSS share basic characteristics with e-coaching systems, in that they too support human operators in complex decision-making processes by providing information and advice in a timely manner. In that respect, there appears to be at least *prima facie* plausibility for supposing parallels between the domains, such that if DSS trigger a decision bias in their operators, we may suppose that the same or a similar mechanism will be triggered by e-coaching systems.

Second, though there certainly are differences between DSS and e-coaching systems, I believe that these differences add rather than subtract to the plausibility that parallels between the domains hold. Let me explain this with two examples. First, consider that one difference between the types of systems concerns the stakes that are involved with the decisions that are being made. In particular, with DSS, the stakes are often high: misdiagnosis in medical domains can have severe consequences for patients, and inattention or inadvertent oversights as an air traffic controller or pilot can have detrimental consequences for the many passengers on board the aircraft or aircrafts being controlled. Because of these high stakes, decision-makers in those domains are trained specifically to be vigilant in their monitoring efforts as well as in their assessment of the information that is provided to them, in

[18] In fact, given the novelty of e-coaching systems, hardly any evaluative studies have been done at all. For exceptions, however, see Klein et al. (2014); Kamphorst et al. (2014a,b); Horsch, Lancee, Griffioen-Both, Spruit, Fitrianie, Neerincx, Beun, and Brinkman (2017).

[19] For an historical overview of the development of DSS, see Power (2008).

the hope of keeping the number of errors to an absolute minimum. In contrast, with e-coaching systems, though the stakes of attaining behavior change can be high (e.g., losing weight may be medically necessary), the stakes of individual decisions are typically low (e.g., consider the decision to drink a glass of water instead of soda). If we pair this with a lack of explicit training to be vigilant, it seems reasonable to suppose that people who use e-coaching systems will typically be more likely to let their guard down with regard to being appropriately vigilant than the professionals using DSS. So, if the evidence shows that DSS trigger a mechanism that leads to a decrease in vigilance, it is reasonable to suppose that the same will hold for e-coaching systems.

Similar reasoning holds for the second example as well. As I stated in Chapter 2, e-coaching systems utilize persuasive strategies to convince people to follow the system's advice. With DSS, this was never the aim. Rather, DSS are designed to maximize the likelihood that important information is brought to the attention of the human operator, to use at his or her discretion in the decision-making process. As I see it, that also means that if we find a decision bias in DSS, where no intentional effort was made to make the system persuasive, we are likely to find it also in systems that do utilize persuasive strategies.

Clearly, none of this is strictly irrefutable; ultimately, empirical evidence will tell whether and to what extent these suggested parallels hold. But I think I have made them plausible enough to review the existing literature on automation-related decision biases in the domain of DSS in order to provide an empirical foundation for the HRC Claim.

4.2.1 Empirical Evidence

In the literature on decision support systems there is evidence that automated decision support, while generally effective in improving performance on a particular aspect of a task (e.g., Garg, Adhikari, McDonald, Rosas-Arellano, Devereaux, Beyene, Sam, and Haynes, 2005), can also lead people to make mistakes. As Skitka, Mosier, and Burdick (1999) put it, "the presence of automated decision aids might reduce one class of errors (those that the automated decision aid has been explicitly programmed to detect and make recommendations about under normal functioning conditions), but [...] they introduce

the possibility of making new kinds of errors" (Skitka et al., 1999, p. 992). These new kinds of errors can be classified into two groups, viz. *errors of omission* and *errors of commission*. As Skitka et al. explain:

> Errors of omission result when decision-makers do not take an appropriate action, despite non-automated indications of problems, because they were not informed of an imminent system failure or problem by an automated decision aid. Errors of commission occur when decision-makers follow automated information or directives, even in the face of more valid or reliable indicators suggesting that the automated aid is not recommending a proper course of action. (Skitka et al., 1999, p. 993)

Research has shown that both kinds of errors can originate from a lack of effort in assessing the appropriateness or correctness of a decision aid's advice. For example, Parasuraman, Molloy, and Singh (1993) found that people performed substantially worse at detecting automation failures during flight simulations when they were supported by a reliable, but not infallible decision aid (Parasuraman et al., 1993).[20] This result was later replicated by Bagheri and Jamieson (2004). In another flight simulation study, Mosier, Skitka, Heers, and Burdick (1998) found an omission error rate of 55% in pilots who were supported by decision support tools (Mosier et al., 1998). In this study, this could mean for example failing to detect mistakes in altitude clearance or not noticing an incorrectly executed heading change. In addition, they found that 100% of the participants in their study erroneously shut down the engine of the (simulated) aircraft in response to a false message about an engine fire, despite traditional indicators suggesting otherwise.

More recently, Bahner, Hüper, and Manzey (2008) reported that interacting with a decision aid that provides advice for fault diagnosis and management in a process control simulation of the life support systems in a spacecraft also led participants to make both kinds of errors (Bahner et al., 2008). It goes without saying that this could have severe consequences in real life, given how critical the life support system is to the atmospheric conditions in a space cabin. Finally, in

[20] Most if not all of the studies mentioned in this section use decision aids that are reliable but imperfect. Moray, Inagaki, and Itoh have suggested that system reliability has to be 70% or up (Moray et al., 2000), otherwise the effects go away.

yet another domain, similar results were obtained in a study where UK general practitioners had to prescribe in twenty hypothetical primary care scenarios. There, Goddard, Roudsari, and Wyatt (2014) found that "in 5.2% of cases correct answers were switched to incorrect answers after [the general practitioners reviewed] the simulated DSS advice" (Goddard et al., 2014, p. 374).

These studies are just a sample of the empirical evidence available, but taken together, these results already quite clearly suggest that there is something about automation that tends to affect how vigilant people are in either detecting failures of a system to provide advice, or assessing the advice they are given (or both). As Skitka et al. write, "[t]he presence of automated cues appears to diminish the likelihood that decision makers will either put forth the cognitive effort to seek out other diagnostic information or process all available information in cognitively complex ways [...]" and, "[i]n the absence of an automated directive, conversely, people often do nothing, regardless of what other system indices imply should be done" (Skitka et al., 1999, p. 994). These are striking observations, and if it is indeed the case that these are effects of automation itself—and it certainly looks that way—then it is likely that the same sort of effects will be found when people interact with e-coaching systems, with people either not noticing a lack of advice, or not assessing the appropriateness of the advice. I will provide concrete examples of the kinds of consequences this could have for people later in the chapter, but the gist of it is easily grasped: people might miss genuine opportunities to course-correct in their self-regulation if they fail to detect automation failure, and they might be persuaded to do something they would later regret if they did not assess the suggestion for action adequately.

At this point, one might object to the generalization, arguing that the situations to which participants in these studies were exposed are not comparable to daily life. For what most of the cited studies have in common is that they studied multi-task scenarios with high workload, while there is evidence from work by Thackray and Touchstone (1989) that subjects who are only tasked with one monitoring task were as "efficient at monitoring in the presence of automation as they were in its absence" (Parasuraman et al., 1993, p. 3). In response, I would argue that, though daily life might not always quite resemble working in a high-tech cockpit, e-coaching systems will work in a multi-task

environment for much of the time for most people, given that they will operate and intervene while people are working, parenting, making calls, exercising, eating, playing, etc. Therefore, I think it is reasonable to say that the risk that people will make errors of the kind described above is real.

In the literature on automation, different proposals have been put forward about *why* these effects occur. The two most prominent ones have been labeled "automation-induced complacency" and "automation bias". Here, though, is where it becomes conceptually muddled. Historically, the former concept has been used predominantly to explain errors of omission (Parasuraman et al., 1993), which has brought forth the persistent, but false belief that complacency can only cause people to fail in detecting system failures (see, e.g., Wickens, Clegg, Vieane, and Sebok, 2015). The latter concept, on the other hand, has been appealed to widely as a general explanatory cause for both kinds of errors (e.g., Mosier, Dunbar, McDonnell, Skitka, Burdick, and Rosenblatt, 1998; Mosier and Skitka, 1999; Bahner et al., 2008; Goddard et al., 2014; Wickens et al., 2015).[21] In addition, the use of the term complacency, without qualifications, to describe operators whose monitoring performance decreased when using decision aids, has been critiqued for conveying undue value judgment on human experts (e.g., Alberdi et al., 2009, p. 21), particularly in those cases where it was unclear whether the performance had even dropped below a normatively optimal level (Moray and Inagaki, 2000; Moray, 2003). Following these critiques, it appears that the case for complacency has been on the decline. This is unfortunate, because while these critiques do provide grounds for adding nuance, they do not undermine the idea that complacency can be induced by automation, and that this can cause people to make certain kinds of mistakes.

To complicate matters further, a number of researchers have tended to equate (or confuse) cause and effect. For example, Cummings (2004) writes that "automation bias [...] *occurs* when a human decision maker disregards or does not search for contradictory information in light of a computer-generated solution which is accepted as correct" (my italics) (Cummings, 2004, p. 2). This implies that either the bias

[21] Others, like Coiera (2015), take a different tack altogether, and consider the terms synonyms. For example, Coiera writes: "*Automation bias* or *automation-induced complacency* is a very specific bias associated with computerised decision support and monitoring technologies" (Coiera, 2015, p. 418, his italics).

is caused by the same cause as the behavior, or that the automation bias *is* the behavior. In equally confusing fashion, Bagheri and Jamieson (2004) say that complacency "*refers* to the [...] decline of [...] monitoring performance" (my italics) (Bagheri and Jamieson, 2004, p. 54), and Reichenbach, Onnasch, and Manzey (2010) refer to complacency *as* "insufficient monitoring or checking of automated functions" (Reichenbach et al., 2010, p. 374).

I realize that in operationalizing certain constructs, it is sometimes useful or even necessary to focus on a resulting behavior. However, conceptually it is important not to confuse the explanandum with the explanans. This is especially so in cases concerning decision bias, where it can be crucial for questions about culpability to keep the bias conceptually separated from the effects the bias may have, as there may be mediating factors in between. As a case in point, consider how people can have an implicit racial bias without exhibiting racist behavior. Similarly, it is imperative to notice that complacency is a state that causes a certain lack of motivated action, rather than it *being* the lack of motivated action.

I want to make clear that by mentioning these examples I do not mean to discredit any of the authors or their work. Undoubtably, in some of these cases it will just be a matter of phrasing, and nothing more. However, my reason for giving the examples is that, on the whole, they appear to be symptomatic of a deeper underlying issue, namely a misunderstanding of what complacency is.

As it happens, I am not alone in making an assertion of this kind. Parasuraman et al. (1993), Alberdi et al. (2009) and Goddard, Roudsari, and Wyatt (2012) have all pointed out that there is no general consensus in the field about what complacency is exactly.[22] Surprisingly, though, recent research efforts have gone towards devising an integrative model that focuses on the shared commonalities between the constructs of automation bias and complacency (see Parasuraman and Manzey, 2010; Kidwell, Miller, and Parasuraman, 2014), rather than providing conceptual clarification. This is unfortunate, as I believe that complacency is a very subtle, but real phenomenon worth thinking about.

[22] See Parasuraman et al. (1993, p. 2), Alberdi et al. (2009, p. 4) and Goddard et al. (2012, pp. 121–122).

So, in what follows, I would like to draw on work from Jason Kawall (2006) to give a more comprehensive and technical account of complacency than I have used thus far, in order to shed light on how I think complacency stands in relation to decision bias. For the present purposes of this dissertation, this will help us to get clearer on what exactly the complacency concern amounts to. Hopefully, in addition, my reflections on this topic will also prove to be helpful to the field of decision support systems.

4.2.2 Complacency Defined

In Section 3.3 I introduced the notion of complacency in colloquial terms of undeserved or unwarranted self-satisfaction that leads to a lack of action or effort. Characterized in this way, the notion of complacency fits well within the broader web of concepts that we have been discussing thus far, including self-regulation failure, responsibility, trust versus vigilance, automation bias and reliability, agency, and self-governance. When searching for a more precise characterization of complacency within philosophy, however, it turned out that complacency is typically discussed in contexts of morality (e.g., Unwin, 1985; Crisp and Cowton, 1994), where it is portrayed as a vice, just like apathy, resignation, pride and hypocrisy.[23]

Since the literature on the concept is dominated by discussions around moral issues, it is especially important to establish that complacency is not limited to issues of morality. To illustrate this, consider the following case description.

> [I]magine a professor who has become complacent about a course she has taught for many years—she no longer reads new material relevant to the course topic, does not change the assignments, and so on, as she believes (culpably) that the course is still good enough as it is. [...] [T]he complacent professor feels self-satisfied as she culpably believes that she has already done enough, that no further work is needed. (Kawall, 2006, p. 345)

[23] Unlike many other vices, however, complacency is often not easily recognizable, as "[c]omplacency does not cause evil or mediocrity; it is a vice that allows these to exist" (Kawall, 2006, p. 343).

On the assumption that teaching a course at university is a non-moral project (though questions of morality may certainly arise when teaching), we can see that complacency is not always associated with doing something morally wrong.

Kawall's citation also highlights two important aspects of the phenomenon itself. First, it makes clear that complacency involves an overestimation of "one's own positive status in an epistemically culpable fashion" (Kawall, 2006, p. 343), which then leads to a state of excessive self-satisfaction. This epistemic culpability comes down to the idea of an agent making irresponsible, faulty judgments about either the demands that are placed upon the agent, or about how well the agent's efforts or actions satisfy these demands. The complacent professor's case is ambivalent in this respect. It could be that the professor *underestimates* the demands of teaching the course, perhaps as result of the demands having changed over the years (something the professor knows or ought to have known). However, it could also be that the professor *overestimates* how much her preparations in previous years count towards meeting the demands of teaching the course this year.

Second, Kawall's case of the complacent professor helps recognize that complacency is scoped, in the sense that one can be complacent with respect to some goods or some projects or some processes, but not with respect to others. The professor thus need not be complacent with regard to her PhD supervisory duties or in fact in any other domain of her life at all.

These two aspects are also present in Kawall's definition of complacency, which is as follows.

> **Complacency (with respect to some good or project G)**: is constituted by **(i)** an epistemically culpable overestimate of one's accomplishments or status that produces **(ii)** an excessive self-satisfaction that produces **(iii)** an insufficiently strong desire or felt need to maintain (or improve to) an appropriate level of accomplishment, that in turn produces **(iv)** a problematic lack of appropriately motivated, appropriate action or effort. (My emphasis.) (Kawall, 2006, p. 345)

This definition brings out another important aspect of complacency, namely its relation to motivation. Herein lies the principle difference between complacency and apathy, a state that is also associated with a

problematic lack of appropriately motivated, appropriate action or effort. However, "[t]he complacent lack of action ultimately arises out of self-satisfaction and a lack of attention; the apathetic out of a complete lack of concern" (Kawall, 2006, p. 350). In the complacent professor's case, it is not that she does not care about teaching the course, but simply that she culpably judges that she has already done enough and that no further effort or action is required of her.

We may suspect that there will be an interplay between apathy and complacency sometimes, in that a lack of concern can "manifest[...] itself in careless, overly generous epistemic assessments of oneself" (Kawall, 2006, p. 348). For example, in a different scenario, we could imagine a professor not caring (or not caring enough) about some internal report, so that she judges that her writing one paragraph on a piece of scrap paper is having done enough (while she ought to know that more will be required of her). Despite this possible interplay, however, we should be mindful not to confuse apathy with complacency or see the two as inseparable. What should be leading is the idea that culpably judging that one has done enough already can produce a lack of effort or action that poses a kind of problem. And part of our project will be to assess the normative implications of the particular problem that complacency poses in relation to people's practical reasoning.

There is a final aspect of complacency that I would like to draw attention to, one that remains underexposed in Kawall's definition. It is about how the concept is related to time. Sometimes, it will be apt to diagnose a lack of effort *at a time* as a result of one's one-time (culpable) lapse in judgement. For example, we might want to say of a professor who, just this once, failed to adequately prepare a lecture, that her failure was due to her being complacent, as she, just this once, culpably overestimated how well she remembered the lecture from the previous time she gave it. And Kawall's definition gives us the vocabulary to adequately express cases like these. More often, however, complacency towards some G will be enduring over a period of time, and this is contingent on the temporally extended nature of our agency. That is, we can expect that unless something significantly changes in one's constellation of beliefs, desires and plans, these states will reliably lead to the same culpable overestimation of one's accomplishments. And, in turn, this overestimation will reliably lead to a problematic lack of effort or action. So, once again, the temporal aspect helps to see why

we would think, normatively, that complacency is undesirable: it is not about a single lapse in judgment, but rather about a routine failure to "get things right".

Crucially, getting things right, and producing the kind of change necessary for breaching complacency, will involve critical self-reflection. This not only involves assessing the accuracy of one's beliefs, but also assessing the coherence and consistency of one's plans and intentions, and reconsidering one's plans and intentions when appropriate. If, however, the scaffolding that is offered to an agent by an e-coaching system facilitates complacency in relation to exactly those processes—and this is really what lies at the heart of the HRC Claim—then this compromises self-governance.

In Section 4.3 I will argue for this implication. First, however, let me elaborate on how I think complacency relates to the idea of an automation-related decision bias, in order to make the case for the HRC Claim.

4.2.3 Automation-Related Decision Bias Revisited

Recall that the HRC Claim posits there is some mechanism that can cause people who use e-coaching systems to become complacent in their practical reasoning. So far, I have attempted to add plausibility to this claim by reviewing empirical evidence that suggests that there are indeed mechanisms that are related specifically to automation which can cause people to decrease their vigilance for a particular task. By itself, however, this does not get us to the HRC Claim quite yet, as I have not explained how the causal mechanism actually heightens the risk of complacency. Moreover, I have not said much about what the mechanism could be, except that I think it is a type of decision bias. In this section, I aim to remedy both shortcomings, beginning with the latter.

Researchers who study automation bias typically subscribe to the idea that people are *cognitive misers* (Fiske and Taylor, 1991), meaning that people use heuristics and cognitive shortcuts to be economically prudent in their reasoning (e.g., Mosier et al., 1998; Mosier and Skitka, 1999; Goddard et al., 2012). The thinking here is that using heuristics allows agents to be effective in their pursuits, given their cognitive

limitations (see also Dennett (1987)).[24] It has been suggested that the bias that people appear to have with respect to automation can be brought under this general heading, and I think this is right. Generally speaking, if someone or something else can reliably take over some task of ours, then that allows us to focus our attention on something else (and this is especially true in multi-task situations). My hypothesis then is that the relevant bias is a hardwired tendency to equate high percentages of reliability above some relevant threshold value with a positive binary value of reliability (i.e., "this system is reliable, full stop"). Often, such cognitive shortcuts will be unproblematic and even helpful, given that trade-offs in attention allocation have to be made and that in order to make those trade-offs it is sometimes necessary to reduce complexities about uncertainties in one's reasoning. In this sense, we can understand the point by Moray (2003) that it is overly hasty to call any shift in attention and subsequent decrease of vigilance an instance of complacency. However, sometimes the task at hand is of such importance that we cannot allow our vigilance to drop below a certain critical level. It is in those cases that we have to suppress our tendency for reliability overestimation.

Here, then, do we finally see how the bias is related to complacency. We may all have a tendency to overestimate the reliability of reliable but imperfect systems, but we become complacent when we go along with this tendency *and thereby* make the epistemically culpable overestimation of how well we have met certain responsibilities or demands by relying on the external system, such that this overestimation produces a problematic lack of appropriately motivated, appropriate action or effort. For example, if one culpably overestimates the reliability of a traffic control system and decreases one's monitoring vigilance below a critical threshold because of this overestimation, knowing that monitoring is one's primary objective, then errors of commission and omission may be attributed to complacency. On the other hand, if errors of commission and omission are due to an appropriate decrease of monitoring vigilance (e.g., if vigilance for another important task increased), these errors may be explained by an appeal to one's bias, but not to one's complacency. So, what this example illustrates, is that it is one thing to trace back errors of omission or

[24]This idea also fits well with the planning theory of agency, which holds that the evolutionary purpose of planning is to work around our cognitive limits. See, for example, Bratman (2007, p. 53).

commission to our tendency to overestimate reliability—this might be the result of using heuristics to make effective use of resources given multiple tasks—and quite another to trace them to a problematic overestimation of our own accomplishments in relation to certain demands that is epistemically culpable because there is an awareness of both the imperfect reliability of the system and the importance of vigilance for a particular task.

Now, what is special about the e-coaching systems case, and what explains why all of this is of concern ethically, is that the tasks for which vigilance might be decreased as a result of complacency are fundamental to self-governing agency. That is, part of being a self-governing agent, as we have seen in Section 4.1, is the ongoing process of striving for a constellation of values, goals, beliefs, plans and intentions that is a coherent and consistent whole. Part of this effort is assessing how well suggestions for action (whether they are internally or externally generated) fit within this constellation. The risk with e-coaching systems, then, particularly with those that are part of the kind of socio-technical systems as described in Section 2.4, is that they provide such a smooth, fluid, reliable user experience combined with expert knowledge that they might lull people into a state in which they overestimate how much they themselves have accomplished by employing the e-coaching system with respect to making sure that one's actions are in line with and because of one's own values. If this is so, and the HRC Claim holds, then this has implications for self-governance. Or so I will argue.

4.3 Implications for Self-Governance

If the HRC Claim holds, then we have established the first part of the complacency concern, namely that e-coaching systems indeed run the risk of facilitating a kind of complacency in their users that affects their vigilance with regard to certain processes related to practical reasoning. But why think of this risk as a concern to mitigate? Why not instead think of this risk as acceptable potential "collateral damage" that is outweighed by the benefits of an increased ease of "staying on course" and having to think less in a world that is already cognitively demanding? The answer lies in a second claim about the implications of such complacency, namely that complacency in processes related

to practical reasoning undermines self-governance. This claim can be formulated as follows.

The Implication for Self-Governance (ISG) Claim
Complacency in practical reasoning undermines self-governance.

Contrary to the HRC claim, the ISG Claim makes a conceptual point about the concept of self-governance. In particular, it highlights an important, but sometimes overlooked aspect of self-governance, namely the appropriate level of vigilance that is needed in the ongoing process of ensuring that one's values, goals, beliefs, intentions and plans together form a consistent and coherent whole. As such, it is a general claim that is not specific to just one account of self-governing agency. That said, the planning model of agency does provide a good conceptual framework for explaining how complacency may manifest itself in one's practical reasoning.

In this section, I will elaborate on and defend the ISG Claim. Subsequently, I will tease apart the different ways complacency may manifest itself in people's practical reasoning. Finally, in Section 4.4, I will give a series of fictive cases that aim to illustrate how these different ways of being complacent can lead to scenarios in which people in the end regret not having been more vigilant. I will discuss these cases using constructs from the planning theory of agency.

4.3.1 The Role of Vigilance in Self-Governance

As mentioned in the Introduction, and repeated in Section 4.1, one key constitutive aspect of self-governance is engaging in a process of determining where one stands in relation to certain influences. Importantly, it matters very little whether these influences are generated externally, or whether they come from within: self-governing agents have to take a stance in relation to, for example, suggestions from other people or from e-coaching systems in the same way that they have to take a stance in relation to their own first-order desires. For example, one might endorse one's desire to help a friend, and reject one's desire to drink a can of paint. By doing so, self-governing agents actively shape their practical standpoint.

That forming a practical standpoint requires *active* engagement by the agent is an important observation. As Anderson puts it, "[a]utonomous persons are active rather than passive with regard to [...] influ-

ences; upon reflection they can endorse or reject influences, thereby taking ownership of them or disowning them" (Anderson, 2013, p. 6).[25]

I want to be careful here to avoid charges of hyper-rationalism (see, e.g., Anderson, 2013, p. 7). Some people might find that talk of active and explicit endorsement is illiberal with regard to self-governance because explicit and critical self-examination is most commonly found in people who are highly educated. The worry then is that an account of self-governance that leans heavily on our capacity for critical thought for endorsing or rejecting influences might limit the self-governance of individuals who live their lives relying more on their emotions and intuitions.

To alleviate this worry, let me be clear that by stating that self-governance requires taking a stance in relation to influences, I do not mean to suggest that we consciously have to deliberate about the possible implications of each influence that affects us. Such a position is quite evidently overly strong. Rather, the way I see it, and given the nature of the planning model, we can think of reflective endorsement or rejection as emerging from an interplay between thought, emotion, and intuition (cf. Bratman, 2007, p. 36).

Perhaps there are other, more nuanced ways of describing this process of endorsing and rejecting influences so that is more immediately evident that the view accommodates individuals who live their lives relying more on their emotions and intuitions than their capacity for rational thought. Depending of course on the precise articulation, I think that such more nuanced views would be broadly compatible with what I am saying here, which is only that as self-governing agents, we may not have full control over the influences that affect us, but we can (and must) claim a certain ownership over some of them. That means that, in order to count as self-governing agents, also those individuals who rely more heavily on emotions and intuition to give direction to their lives have to take an active role in determining who they are by identifying with some influences, and not with others.

Understanding forming a practical standpoint as an active rather than a passive process helps to see that self-governance *requires* a certain watchfulness and diligence with regard to assessing influences.

[25] Notice that in a theory of agency such as the planning theory, this talk of agents standing back from and reflecting on their attitudes is shorthand for the more intricate model of hierarchical structures of reflexive mental states that have agential authority in light of their functionings. For a discussion, see Bratman (2007, p. 196).

This watchfulness and diligence corresponds to the idea that I have been using throughout the dissertation, namely that there is a minimal level of vigilance that self-governing agents need to keep up. If we are overly lenient, and let our vigilance drop, we undermine our self-governance by allowing inconsistencies to be introduced into our constellation of beliefs and intentions and similar mental states, and by missing opportunities to correct such inconsistencies.

Notice that this undermining should not be thought of in terms of risks. As I mentioned in the beginning, the ISG Claim concerns a conceptual implication, in that *if* one becomes complacent in aspects of one's practical reasoning, *then*, by necessity, one compromises one's self-governance, regardless of whether any inconsistencies or incoherences have yet been introduced into the constellation of one's mental states as a result. The extent to which one's self-governance will be compromised will depend on the particulars of the circumstances and will therefore vary, but the implication thus holds regardless of the consequences of one's complacency. In contrast, experiencing negative consequences from a lack of vigilance *can* be expressed in terms of risks. That is, it would be appropriate to say that complacency increases the risk of people making certain kinds of errors that can be associated with decreased self-governance.

To see what these errors might consist in, it will be instructive to revisit the conceptualization of self-governance from Section 4.1. Before we do so, however, there is one objection to the ISG Claim that needs to be rebutted before we can move on. That objection pertains to the extended theses as discussed in Section 2.4, and asks the question why we should take the biological human being as the relevant unit of interest when it comes to self-governance instead of the socio-technical system of human being plus e-coaching system. I will address this issue in the following subsection.

4.3.2 Objection: Extended Self-Governance

The ISG Claim makes a statement about the relation between complacency in one's practical reasoning and one's self-governance. To be more precise, I have argued that self-governance is undermined when an epistemically culpable overestimation of one's accomplishments produces a problematic lack of effort in keeping up vigilance

in the process of assessing influences.[26] The implicit assumption in this argument has been that the "self" in "self-governance" refers to a body-bound entity who is demarcated from the world by the skin-skull boundary. However, in line with what we have seen in Section 2.4, this perspective may be challenged on grounds that it is sometimes appropriate to think of concepts such as cognition and the will as extending beyond the skin-skull boundary and into the world. Proponents of extended theses, might therefore question the implicit perspective that I have taken about the self, and argue that if we were to consider the self as extended, then reduced vigilance in the body-bound entity need not imply an undermining of the self-governance of the socio-technical system as a whole. After all, they might say, it is conceptually possible that the level of vigilance within the extended system is kept the same, even if there has been a shift in *where in the system* the assessment of influences takes place. Supposing that this is the case, what reason, other than an appeal to an internalist bias, do we have to think that the reduced vigilance in the body-bound entity is problematic for self-governance?

I see two ways of addressing this objection. The first approach is to charge the opponent with changing the rules of the game mid-play by shifting perspective only in relation to self-governance and not in relation to complacency. With this approach, one could concede that there are potential situations in which a lack of effort in keeping up vigilance in the body-bound entity is indeed unproblematic, as long as we make clear that those situations are *therefore* not cases of complacency. That is, because the ISG Claim presupposes complacency, and complacency involves, by definition, a *problematic* lack of effort, the starting point should always be that there *is* a problematic lack of effort with regard to vigilance produced by an epistemically culpable overestimation of one's accomplishments, regardless of the perspective we take. So, if we do take an extended perspective and think of the self as referring to the socio-technical system, rather than to the body-bound entity, then a fair reading of the ISG Claim requires that we assume a problematic lack of effort in the vigilance of the extended system, in which case the implication of the ISG Claim goes through normally.

[26] Notice that I am skipping intermediate steps (ii) and (iii) from Kawall's definition here for the sake of brevity.

While I think this first approach results in a valid rebuttal, I find it not quite satisfactory as it asks us to concede too much, too quickly. After all, it requires adopting a view of agency that is so very fluid that it allows for processes involved with determining our practical standpoint—that by which we shape our identities—to be performed, at least in part, by an external system with which we have coupled. Putting aside the obvious practical problems with this—it would seem that the system would need direct neurological access to the contents of our mental states for example—do we really want to make this concession?

I think we can avoid the concession by resisting the objection in another way, even while still acknowledging the strengths of extended theses (as I have done in Chapter 2). This second approach revolves around the observation that neither the extended mind or the extended will implies extended agency. That is, we can allow agents to have the ability to interface with the world and form extended systems, without thinking of the agent itself as being extended. Rather, the picture is that the agent plays an important role within the extended system, but is not identified with the extended system. This position is what Wilson (2004) has described as the *narrow subjects, extended systems* view (Wilson, 2004). On this view, the agent is demarcated by its body, where the agent's locus of control is. As Wilson argues, this view fits well with our ordinary understanding of the world, and parallels the way we typically think about acting: the agent is body-bound, but the agent's actions extend into the world. Similarly, then, we can hold that the agent is body-bound, but that the agent's cognition as well as the agent's will extends into the world.

Once we acknowledge that the agent is body-bound, we can see that the implicit perspective was the more appropriate perspective to take, given how closely self-governance is tied to agency. With "narrow agents", whose bodies are "the immediate locus of willed action [and] the gateway to intelligent offloading" (Clark, 2011, p. 207), there are processes that cannot be outsourced without changing the very being of the agent. In particular, if agents are body-bound, then the processes that govern the outsourcing of certain other tasks or processes, as well as the processes that are involved with the agent's practical standpoint, also have to be body-bound. Often, I suspect, these processes will overlap, but the reasons for thinking that these processes should be

body-bound are subtly different.

The first reason concerns the responsibilities that accompany delegation. When we delegate a process or a task, it is important to realize that, though we outsource the execution as well as the responsibility for execution, we do not abdicate our accountability, and therefore retain a responsibility to assess the quality of the execution, and to check whether it was performed within the pre-determined boundaries that we set beforehand. *These* processes, then, cannot themselves be outsourced to an external system.

As an example, consider again the original memory-based case from Clark and Chalmers (see again Section 2.4). About Otto and his notebook, Clark (2011) writes that "any information [...] retrieved [from the notebook should] be more or less automatically endorsed" and "should not usually be subject to critical scrutiny" (Clark, 2011, p. 79). This careful formulation indicates delegation, as the (body-bound) agent has to, however minimally, assess and endorse the information from the notebook. This means that Otto will typically use the information from the notebook without issue—for example to find his way to museums—but also that Otto would notice and hesitate if the handwriting in the notebook did not look as his own.

In Otto's case, the process of assessing the information is quite minimal and would therefore not be effortful. Insofar that e-coaching systems offer reminders of previously made plans, we may surmise that agents who use such systems can also suffice with these minimal endorsement requirements. Notice, however, that this minimal level of scrutiny is only appropriate *given* that "the information has been consciously endorsed at some point in the past and indeed is there as a consequence of this endorsement" (Clark, 2011, p. 79).[27] So, when e-coaching systems make new suggestions, then we have to assume that the rules are different, and that more careful scrutiny is required of the agent.

The second reason we have for asserting that certain processes have to occur within the skin-skull boundary is that these processes play a constitutive role in the agent's self-governance. Provided that agents are indeed body-bound, then it follows that for an agent to be truly self-governing, the processes associated with "laying down the law for

[27]For this reason, according to Clark, access to Google on one's mobile phone for example does not count as part of one's cognition. See Clark (2011, p. 80).

oneself" should also be body-bound. In other words, though external systems—or other people for that matter—may offer suggestions for action, it is the agent who has to endorse or reject the suggestion and form the appropriate corresponding mental state, thereby contributing to the agent's practical standpoint.

Granted, one could argue that there are (or can be) extended agents, and that for those agents the process of endorsing or rejecting influences *can* be performed by an external resource if the parity principle holds, as that external resource in that case can be considered as *part of* the extended agent. However, I see no compelling reason to think that human beings already are such extended agents, even if their will and their cognition sometimes extends beyond the skin-skull boundary. I find that the narrow subject view offers a reasonable metaphysics of agency, and I think the onus of proof is on the extended agency proponent to show otherwise.

With this, I have hopefully done enough to establish some general plausibility for the ISG Claim. So far, however, I have not been very specific about the different ways in which one's complacency in one's practical reasoning may manifest itself, nor have I been concrete about the consequences this may potentially have. Therefore, in the remainder of this chapter, I will use constructs from planning theory to distinguish two main ways in which one's complacency may manifest itself, and provide fictive cases that lend further support for the ISG Claim and, in addition, illustrate the sort of consequences that these different facets of complacency in one's practical reasoning may have.

4.4 Illustrative Examples

In Section 4.1 I explained that on the planning model, self-governance is conceptualized as being guided in one's behavior by reflexive valuings that have agential authority and with which the agent is satisfied. In this formulation we see that self-governance has more stringent conditions than planning agency itself. As Bratman himself points out, "one's planning agency may be tied to the pursuit of ends that are compulsive or obsessive or unreflective or thoughtless or conflicted in ways incompatible with self-government" (Bratman, 2007, p. 198). Self-governance, in contrast, requires that one's ends are freely chosen

and that one is satisfied with, and thereby not estranged from, one's higher-order policies. Moreover, it is required that one's behavior is *guided* by these policies "in the right way". And while I suggested that the causal chain from one's valuings to one's actions can, in principle, involve an external system, it is important that one does not let the external system steer one's behavior in a direction that is incompatible with one's valuings. This means that for an agent to be self-governing on the planning model of agency, the agent has to be vigilant in assessing an e-coaching system's suggestion for means-end coherence, as well as for consistency.

In what follows, I will present a set of fictive cases that illustrate different ways in which complacency may manifest itself, and the consequences this may have. Note that these cases are stylized to make specific points, and should not be taken as an expression of an expectancy about outcomes of the prototypical use of e-coaching systems. After all, not everyone will be susceptible to the effects of automation bias to an equal degree, and it is good to remember that certain technology-savvy users might have much more nuanced views about the limits of certain sensor systems and corresponding measurements.[28] Moreover, different e-coaching systems will also differ in their design, and therefore perhaps also in their tendency to facilitate complacency (see also Section 5.2). Still, given what we have seen in Section 4.2.1, it is plausible that the kind of complacency I am concerned with can potentially occur with a subset of the users of e-coaching systems. Since there are no actual case descriptions of such behavior available yet given the general state of development of e-coaching technologies, I will present fictive cases that are meant as illustrations to what I have been arguing in this chapter. With that disclaimer in place, let me briefly say what binds the cases together, and what sets them apart.

At the basis of each case, we have a person who desires to improve a certain aspect of his or her life, who employs an e-coaching system for this purpose, but who is in some way culpably naive about the extent to which employing the e-coaching system relieves him or her from demands of self-governing planning agency. In order to focus more on the different aspects of these demands, the individual cases

[28] Interestingly, though, a recent study suggested that gamers, who are typically fairly technology-savvy, actually had a higher propensity to overestimate automation reliability (Clare, Cummings, and Repenning, 2015).

do not elaborate too much on this naiveté, so I will make some general remarks about it here.

The idea is to present cases in which the agents in question complacently follow advice from an e-coaching system without being sufficiently vigilant in assessing the advice for coherence and consistency. The problematic nature of this lack of vigilance is most easily illustrated by scenarios in which complacently following the e-coaching system's suggestion leads to negative consequences for the agent, so that will be my starting point. However, as I will argue, negative consequences should not be considered a necessary condition for diminished self-governance, by which I mean that the ISG Claim also holds if the complacent following of the advice turns out to be harmless or even beneficial to the agent.

Key to illustrating complacency in each of the cases will be that the agents observe aspects of the environment and of themselves that are pertinent to an e-coaching system's suggestion, but culpably do not bring these observations to bear, in an appropriate way, on the question of whether or not to endorse and follow that suggestion. In this way, the cases mirror the kind of behavior demonstrated by participants in the empirical studies discussed in Section 4.2.1 where relevant, observed clues from traditional indicators would be culpably dismissed in favor of the decision support system's (mistaken) reports (e.g., Mosier et al., 1998). Thus, though each case may certainly have variants in which there is either no lack of vigilance, or in which the lack of vigilance is excusable (e.g., because of a high-pressure situation that requires the agent to make cognitive trade-offs), the guiding thought is that in the cases demonstrating complacency a moment's reflection would have been sufficient to adequately assess the e-coaching system's suggestion. With those remarks in place, let us examine the cases.

4.4.1 Neglecting to Assess Suggestions

Let us begin by considering three cases that appear to be relatively straightforward illustrations of the ISG Claim. In each of these cases, I am assuming that the agent in question is a planning agent.

A. Alice employs an e-coaching system for learning to deal with anger management issues and follows the system's advice to jump out of a nearby window.

B. Bernard employs an e-coaching system as part of a sleep-restriction therapy program to help with his insomnia and follows the system's advice to go to bed at 9pm instead of his usual, physician-recommended bedtime of 12pm despite many failed attempts in the past to fall asleep at an earlier hour.

C. Chanda, a practicing Muslim, employs an e-coaching system for promoting weight loss and follows the system's advice to have a beef salad for lunch at a non-Islamic restaurant.

In light of the Heightened Risk of Complacency Claim (see Section 4.2), I take there to be a *defeasible presumption of automation-related complacency* in each of these cases, in the sense that each of these agents appears negligent with regard to assessing the e-coaching system's suggestion for coherence and consistency, such that they would have acted otherwise if they had been more vigilant and less trusting on the e-coaching system to make fitting suggestions. As I stipulated that all three individuals are planning agents, intuition suggests that if only they had taken care to assess the e-coaching system's suggestion with regard to their respective constellations of values, beliefs and plans, Chanda would surely have concluded that the beef at the non-Islamic restaurant would most likely not be halal and would therefore not a good option for her in light of her faith, Bernard would have surely concluded that the system's recommendation was conflicting with his knowledge about himself and his condition, and Alice would have surely concluded that the suggestion to jump out of a window was most likely due to a glitch and that actually following the suggestion would not be in her best interest.

Temporarily putting aside the defeasibility of the presumption of complacency, and supposing for the moment that the agents in cases A–C are indeed complacent with regard to assessing their e-coaching system's suggestion, it follows from the ISG Claim that the agents' synchronic self-governance is diminished. And this is plausible, given that the agents more or less blindly do as they are told (this is most blatantly evident in case A.). Importantly, that their synchronic self-governance is diminished is so regardless of the consequences of their respective actions: even if Alice were to land on her feet, Bernard would have a good night's sleep and Chanda would be served halal beef

by an attentive chef, the evaluation of the actions as non-self-governed actions should remain the same. What matters for self-governance, after all, is whether an action is caused and guided in the right way by the agent's valuings, which does not appear to be the case for the agents in A–C.

Let us now take a closer look at the question of whether the agents in cases A–C are indeed complacent. The reason for speaking of a defeasible presumption of automation-related complacency is that, unless it would have been explicitly stipulated, it is not evident from the case descriptions alone that

a) there is indeed a problematic lack of effort with regard to vigilance, and

b) that the problematic lack of effort, if it exists, involves excessive self-satisfaction that can be traced to an epistemically culpable overestimation of the agent's accomplishments (see again Kawall's definition in Section 4.2.2).

For example, consider the following variants of scenario C:

C′ Chanda, a practicing Muslim, employs an e-coaching system for promoting weight loss and follows the system's advice to have a beef salad for lunch at a non-Islamic restaurant, *after having accepted the system's reasoning in favor of a lean and protein-rich meal*.

C″ Chanda, a practicing Muslim, employs an e-coaching system for promoting weight loss and follows the system's advice to have a beef salad for lunch at a non-Islamic restaurant *after having endorsed the system's reasoning in favor of a lean and protein-rich meal in light of a false belief that this particular restaurant happens to only serve halal beef*.

C‴ Chanda, a practicing Muslim, employs an e-coaching system for promoting weight loss and follows the system's advice to have a beef salad for lunch at a non-Islamic restaurant, *after having endorsed the system's reasoning in favor of a lean and protein-rich meal, and having further concluded that this could well mean having a meal that, just this once, was not halal*.

In these variants of scenario C, it is clearer than in the original formulation that Chanda was not completely subservient with respect to the system's suggestion, but it is not obvious at which point she would have been vigilant enough for the presumption of complacency to be defeated. For example, in C′ she is vigilant to the extent that she checks the suggestion for coherence, but she does not notice the inconsistency of her intention to eat non-halal beef with her (faith-related) values. Was she vigilant enough (and perhaps unlucky) or was she complacent?

This brings to the foreground two important but difficult methodological questions about assessing complacency. The first question concerns the standard of vigilance against which to measure an agent's efforts. Intuitively, it is reasonable to expect a higher level of vigilance from agents in certain scenarios than in others; for example, it seems reasonable to expect more vigilance from someone receiving suggestions about medical treatments than from someone like Chanda who is simply ordering lunch. But how should such a context-dependent evaluative standard for vigilance be determined?

The second question concerns determining the culpability of agents given a certain standard of vigilance. Turn now to case A: given some standard of vigilance V (however determined), intuition suggests that Alice falls well below V by failing to assess (and by following) an evidently absurd suggestion. However, even if this is so, there is still the further question of whether we should say that Alice was culpable for this lack of vigilance. If Alice overestimated her accomplishments with regard to ensuring coherence and consistency of her intentions in light of having employed a state-of-the-art e-coaching system, then the answer is surely positive. However, what if we were to learn that at the time of the suggestion Alice was in a delirium caused by new medications she was prescribed? In this variation of scenario A (call it scenario A′), it would not be unreasonable to excuse Alice for falling below V.

A similar point can be constructed for case B as well: given some standard of vigilance V (however determined), intuition suggests that Bernard falls short of meeting V by not (adequately) assessing the newly suggested bedtime in light of his previous experiences with going to bed earlier and the recommendations made by his physician. Still, there is the further question of culpability: did Bernard culpably

think he had done enough with regard to ensuring coherence and consistency of his intentions by employing a personalized e-coaching system, or did he just happen to have received shocking news about a friend's passing (call this scenario B/) and might he for that reason be excused? These are complicated matters to adjudicate and doing so will require a great deal of information about the circumstances of each individual case.

As a means of approaching such matters in a systematic way, I propose to utilize a "reasonable person standard" here as found for example in contract law, criminal law and civil rights law (cf. Schmidt, 2007). Although certainly not without its own challenges (e.g., see Donovan and Wildman (1980); Moran (2003)), a reasonable personal standard could provide a way of determining a fitting standard of vigilance given a certain context by asking what a reasonable person would do in the relevant situation.[29] Then, once the standard is determined, it can be assessed whether the individual in question did or did not meet this standard. Finally, as a last step, if the individual is found to fall short of the reasonable personal standard, it can be examined whether there are mitigating or excusable circumstances to take into account with regard to the question of culpability such as an agent's (temporarily) diminished capacities.

To see how this might work, consider once again case A. First, we ask what a reasonable person would have done in that situation. Here, a plausible answer is that a reasonable person would have noticed that jumping out of a nearby window was neither means-end coherent in relation to the end of dealing with anger management issues nor consistent with one's values. This then determines the standard for Alice: she ought to have noticed the incoherence and inconsistency of the suggestion. Since she did not, she fell short of the standard. Was she culpable for doing so? Yes, for case A without qualifications; no, potentially, if she was in a medication-induced delirium (A/).

Surely, there is much more to say about such a reasonable person standard, but having made the suggestion, I will leave that for future work. Now, I will conclude the chapter with two more cases that aim to add a touch of realism to the complacency worry by providing a sense of how automation-related complacency may develop over time and the consequences this may have. Adding the temporal dimension helps

[29] Note that the reasonable person is not the same as the average person.

demonstrate how complacency can creep in when using an e-coaching system for a longer period of time, leading to a sustained "problematic lack of appropriately motivated, appropriate action or effort" (Kawall, 2006, p. 345). So, just as Holton's temporal perspective on weakness of will helped reveal the wider range of volitional issues beyond akrasia (see again Section 1.1), the temporal perspective is similarly crucial for understanding the full force of the complacency concern. Moving forward, then, the first case is concerned with automation-related complacency leading to means-end incoherence in one's plans (Section 4.4.2), the second case with complacency leading to plan inconsistency (Section 4.4.3).

4.4.2 Falling Short in Assessing Suggestions for Means-End Coherence

We have seen in Section 4.1 that the planning theory presupposes the existence of rational pressures to which planning agents are subject. One of those pressures concerns means-end coherence, which, to reiterate, gives expression to the idea that it would be pro tanto irrational for a planning agent to intend an end E, believe that M is a necessary means to achieving E and that M requires the agent to now intend M, and yet not intend M now. Recall the agent who plans to visit Ottawa in the fall: having that plan puts the agent under rational pressure to also intend to buy an airplane ticket and to arrange accommodations. This rational requirement of means-end coherence also entails that planning agents have to figure out what a plan demands and what the means to a specific end *are*, especially since these are often not given externally.[30]

It is straightforward to see how e-coaching systems could potentially lighten the burden of this task. For e-coaching systems will contain expert domain knowledge, such that they know—in broad, abstract terms—what is required to achieve a certain end. To take just one example, if one were to employ an e-coaching system designed to help people to go to bed on time, we may assume that the system will have some representation of the intermediate steps necessary to

[30] There are exceptions, like when one plans to do something that has strict formal requirements such as getting a driver's license, but even then should one find out what those requirements are, and how (and when) one is going to fulfill them.

get there. Agents who use such systems can, in principle, make use of this expert domain knowledge in their efforts towards attaining their end. Especially in domains where agents have little experience, or where figuring out means to specific ends will be a complex and time consuming activity, we can see how e-coaching systems could be beneficial. At the same time, regardless of how welcoming an individual may be of these systems, it is important to emphasize again that using an e-coaching systems does not abdicate one from the responsibility to strive for means-end coherence. For, as we have seen in the cases A–C from the previous section, relying excessively on e-coaching systems to the extent of being complacent in relation to critically assessing the system's suggestions is opening the doors to trouble. To further stress the problematic nature of complacency in relation to one's efforts to be means-end coherent, consider now the case of Dave, another planning agent, who not just incidentally follows one wrong suggestion, but who does so consistently over a period of time.

D. Dave is a hard-working professional who experiences high levels of stress, that, among other factors, are related to having an overly demanding agenda. To help him find ways of alleviating some of his stress, Dave employs a well-known, highly recommended e-coaching system designed specifically for stress reduction for working professionals. The system begins by asking Dave about his preferences (e.g., how he would like to be notified, etc.), his hobbies, and about his stress levels, before it makes its first recommendation, namely that Dave performs some breathing exercises during work breaks. Dave complies, and finds the exercises helpful.

   During the second week, the system makes a number of other suggestions as well, including suggestions to meet up with friends and, having learnt that Dave enjoys playing tennis, to play friendly tennis matches. It explains to Dave that both of these activities have been known to reduce stress levels. Dave complies, and over the next six weeks, he continues with the breathing exercises and also makes an effort to go to dinner parties and to play

> tennis regularly. The e-coaching system supports him in these efforts by using Dave's social media accounts to suggest available tennis partners, to recommend suitable meetups that his friends happen to be organizing, and by sending him reminders. During this time, Dave observes that he is becoming increasingly restless at night when he is trying to fall asleep, and that he is shorter than usual with his co-workers. Still, he keeps following the system's suggestions.
>
> After these six weeks, Dave is more stressed than before he employed the e-coaching system. One day, he experiences a panic attack while he is at work. In response, his manager sends him home.

Let us begin by making explicit what exactly the issue is in this case. For Dave, his problem was that though the suggestions to meet up with friends and to play tennis may have seemed (and perhaps were) fitting for stress relief in general, they actually exacerbated stress in his particular case by adding items to his already overly demanding agenda. As a result, Dave had to engage in even more interpersonal planning, leaving him with even less "slack time" than before, which ultimately led him to feel worse.

Like in the cases A–C, this case description also does not explicitly stipulate that the agent in question was complacent. Again, however, given the Heightened Risk of Complacency Claim, I take there to be a defeasible presumption of automation-related complacency, in the sense that, overall, Dave appears to culpably fall short in his efforts to ensure means-end coherence in his plans. Had Dave been more vigilant in that regard, the initial thought is, he surely would have realized that adding social activities to already fully-packed, overly demanding, stressful days was not helping him reduce his stress levels, especially given the self-observed deterioration of his mood and increase of restlessness. A reasonable person standard would, I believe, underline that Dave fell short, since a reasonable person would plausibly have connected the deterioration of his mood and ability to fall asleep to the added social activities and concluded that those activities were not a means to his end of reducing stress.

Note, however, that the temporality of Dave's case adds a dimension

to the complexity of determining this standard. After all, it seems reasonable to expect a different level of vigilance at different points in time. At first, when the initial suggestion is presented, and before there are any actual experiences with following the suggestion for adding social activities to take into account, the level of vigilance required of a reasonable person may be quite low, in the sense that all that may strictly speaking be required is a superficial check that engaging in a social activity may indeed seem like a plausible way of relieving stress (cf. with the absurd suggestion to jump out of a nearby window from case A). Dave may well have met this initial standard of vigilance. However, as times goes by, and evaluative observations are (or ought to be!) made, the standard of vigilance for the reasonable person goes up: a reasonable person is expected to infer from multiple negative experiences that the suggestions made by the e-coaching system are, in fact, not good suggestions for him or her in relation to his or her end, even if the general reasoning behind the suggestion is sensible enough. It is here that Dave's behavior is problematic: though the exact moment at which Dave began falling short of the increasing standard of vigilance may remain vague, it is reasonable to say that Dave fell short from some point in time onwards between the initial suggestion and the suggestion before the panic attack occurred. Let us designate this moment in time as $t_{D<V}$.

Now, was Dave complacent from $t_{D<V}$ onwards, or should he be excused for falling short of the relevant standard of vigilance? The answer will depend on the specifics of the case that, as of yet, are underdetermined. If there are excusable circumstances, for example if Dave's mental capacities were diminished by the chronic stress, then perhaps he should be excused. Barring such circumstances, however, the presumption of automation-related complacency points towards a version of the case in which Dave underestimated what was required of him from a standpoint of self-governing agency in light of him already having employed a highly recommended e-coaching system and having had a good first experience with it (i.e., the breathing exercises) that could well have strengthened the sense of trustworthiness and reliability he felt toward the system. This then would have led him to make an overly generous assessment of his own accomplishments with regard to determining the means to his end, which led to excessive self-satisfaction that resulted in a lack of vigilance with regard to

assessing the e-coaching system's suggestions for means-end coherence. As a result, it would seem that Dave only superficially followed the general reasoning by the e-coaching system instead of reflecting further on whether adding more social activities would benefit him or were benefitting him in his specific case.

Let us assume then that Dave was complacent from $t_{D<V}$ onwards. What does this mean for his self-governance? To answer this question, we must examine synchronic self-governance and diachronic self-governance separately. I will begin with the former.

Given what we know about Dave, it is a fairly safe assumption that Dave is at least capable of self-governance: broadly speaking, there appears to be sufficient unity and organization of the motives of his actions for agential direction, and it seems reasonable to assume that Dave's psychic economy would include self-governing policies (e.g., to give weight to considerations pertaining to his health) for which there were no tendencies or inclinations to change (i.e., that Dave's psychic economy was in a state of satisfaction). Therefore, if, at the moment the first suggestion to meet up with friends was made, Dave believed that meeting his friend could be a means to his end, and then intended to meet his friend in part because of his self-governing policy to give weight to considerations about his health, then Dave's going out to meet his friend seems to be guided in the right way for synchronic self-governance.

However, if we skip ahead to a point in time after $t_{D<V}$ when Dave receives another suggestion to meet up with friends and follows it more or less blindly, without taking into account his evaluative observations from the previous times he followed this suggestion, then Dave's synchronic self-governance is blocked because, like in cases A–C, the ensuing action is not guided in the right way by the relevant practical standpoint. Potentially, however, matters at this point in time have shifted for the worse. For if Dave retained his intended end to reduce stress, and had recognized not just that adding activities was unhelpful, but that freeing up his agenda sooner rather than later would be a necessary means to his end, but did not intend to free up his agenda as he was spurred on by the e-coaching system to keep socially active, then Dave was means-end incoherent. As a result, Dave's practical standpoint itself will have eroded, since incoherence of plans against one's beliefs, according to the planning

theory, "normally baffles the coherence of [a] plan-infused standpoint that is needed for there to be a clear place where the agent stands with respect to relevant issues" (Bratman, 2018, p. 213). If this line of reasoning goes through then Dave's synchronic self-governance at this point in time was undermined not because of a breakdown in the chain from practical standpoint to action, but because there no longer was a coherent practical standpoint.

So how does all of this affect Dave's diachronic self-governance? Given what we know about the outcome of the case, it stands to reason that synchronic self-governance would have been blocked at various later times as well in much the same way. If we recall that diachronic self-governance on the planning theory supposes synchronic self-governance at relevant points in time along the way, and we link that to our conclusion that synchronic self-governance in Dave's case was repeatedly undermined, then it is a straightforward conclusion that Dave's diachronic self-governance was in effect also undermined by the complacency that was brought about in connection to Dave's employment of the e-coaching system. To be sure, Dave was potentially diachronically self-governing in other aspects of his life, but considering that being sent home after a panic attack is clearly not what Dave wants out of life, it is a plausible conclusion that Dave was not diachronically self-governing with respect to dealing with his accruing stress.

This concludes the first illustrative case, which was concerned with the coherence of one's plans in relation to a specific end. As we have seen, however, planning agents are under another rational pressure as well, namely to make sure that their plans are *consistent* overall. To see what happens when people are complacent in regard to critically assessing consistency, let us continue to the final case of Edward, another planning agent.

4.4.3 Falling Short in Assessing Suggestions for Consistency

The planning theory also holds that planning agents are under rational pressure to make sure that their plans fit within their broader constellation of beliefs and plan-like states. This consistency in plans is facilitated by the *filtering role* that plans play in subsequent practical

reasoning. Options for action that are incompatible with existing plans are for a large part simply not generated or quickly dismissed. When options for action are put forward by external sources, however, it is less evident that an agent's plan-like states will fulfill their filtering role. After all, one might act on a suggestion directly (cf. again Alice from case A.), without submitting it first to critical assessment. Doing so might sometimes be tempting, especially when the agent culpably makes a mistaken judgment to have done enough to ensure consistency, for example by providing input to the external source. Relating this more concretely to e-coaching systems, consider now the following case.

E. Edward employs a highly recommended e-coaching system to help him lose twenty pounds over the course of a year. The system starts out by asking Edward about his preferences and subsequently begins collecting data (biometric data, social media data, but also GPS, etc.). The system quickly identifies a number of behavioral patterns that are obviously not benefiting Edward's goal of losing weight. For example, on the basis of a number of photographs that Edward took of the contents of his fridge, the system notices that the fridge is stocked with a wide variety of sugared sodas and beers. The system recommends to Edward that he replaces one soda per day with a glass of water. Edward does so, and as a result, he loses one pound in two weeks time. Some more time goes by, and Edward's progress comes to a halt. Recognizing the lack of progress, the system then suggests to Edward to stop drinking soda altogether. Edward complies, and again starts seeing results.

   At the next weight loss plateau that Edward hits, the system decides to focus on Edward's beer drinking habits. First, it suggests to Edward to stop having beers daily at the end of the night. Edward complies, and as result, he starts losing weight again. Then, five months in, he hits another plateau. Having collected and analyzed Edward's data all this time, the system knows that once a week, on Friday night, Edward goes out for drinks with

> friends, deep into the night. On those occasions—when Edward and his friends recount the respective weeks they have had, and talk about past adventures and future plans—Edward is used to drinking copious amounts of beer. The system reasons that the abundance of alcohol, together with the significant shift in his bedtime, is blocking Edward's progress towards his goal. As a next step, then, to help Edward overcome his current plateau and work towards his long-term weight loss goal, the system suggests that Edward reduce his drinking on Friday nights by staying in rather than going out; a suggestion that the e-coaching system routinely repeats from then on out. Edward complies and as a result begins seeing progress again. Edward feels pangs of regret every time he forgoes his evening out with his friends, but he complies nonetheless, intending to go out with his friends again next Friday. When next Friday comes around, however, Edward again acts on the e-coaching system's suggestion to stay in. Finally, after only eleven months, Edward hits his twenty pound weight loss target.
>
> Content though he is with having fulfilled his goal, Edward is left feeling a sense of emptiness. Reflecting on this feeling, he reaches the conclusion that he regrets not having spent more time with his friends.

Let us again begin by making explicit what exactly the issue is in this case. For Edward, his problem was that though the suggestions to skip his Friday nights out were fitting for him in the sense that they helped him move towards his goal, they at the same time were not congruent with his other plans and values about spending time with friends. Put differently, the e-coaching system's suggested actions were means-end coherent, but they were not consistent with Edward's constellation of beliefs, plans and values. In this regard, Edward's case thus differs from Dave's.

That said, there are also clear similarities between cases D and E. One structural similarity is that Edward, like Dave, acted on suggestions from an e-coaching system while seemingly falling short in his efforts to ensure that the e-coaching system's suggestions were fit-

ting for him in his particular situation. Given the involvement of an e-coaching system, and given the Heightened Risk of Complacency Claim, I again take there to be a defeasible presumption of automation-related complacency. As such, there is again need for a reasonable person standard, and this brings into view a second similarity, namely that Edward's case also has the temporal dimension that is relevant for determining the reasonable standard of vigilance as this standard will plausibly be different at different points in time. After all, there is a difference with regard to the information that is available about the effects of following the suggestion to stay in on Friday night between the first and the last time the suggestion is made.

Initially, we may surmise, the suggestion to stay in on Friday night might seem sensible enough: a reasonable person might plausibly find it a trivial sacrifice to be made this once in the hopes of breaking the current weight loss plateau. However, as time goes by, and evaluative observations are made (i.e. the pangs of regret), the standard of vigilance likely goes up: for a reasonable person will plausibly be expected to infer from multiple negative experiences that the suggestions made by the e-coaching system are, in fact, poor suggestions for him or her in relation to his or her other plans and overall values, even if the suggestions represent means to one of his or her ends. It is here that Edward's behavior is problematic: had he spent more effort in assessing the e-coaching system's suggestions for consistency, he surely would have realized that staying home *this* Friday was not compatible with last week's intending to see his friends next Friday. So, though the exact moment at which Edward began falling short of the increasing standard of vigilance may remain vague, it is reasonable to say that Edward fell short with regard to assessing consistency from some point in time onwards between the initial suggestion and the suggestion before he reached his weight loss goal. Let us designate this moment in time as $t_{E<V}$.

Like in Dave's case, we can now ask whether Edward was complacent from $t_{E<V}$ onwards, or whether he should be excused for falling short of the relevant standard of vigilance. The answer will again depend on specifics of the case that are underdetermined. However, barring excusable circumstances, it is plausible that Edward would have culpably judged to have done enough with regard to ensuring consistency: he had thought about and committed himself to his goal of losing weight,

he had employed a highly-recommended e-coaching system, he had input his preferences, and had even recognized that each suggestion was made in service of his weight loss goal. Having done all of that might have given Edward cause to culpably judge that he had done enough to ensure consistency in his plans, giving rise to excessive self-satisfaction. As a result of his excessive self-satisfaction, he was then insufficiently motivated to engage in the critical assessment—mistakenly assuming that acting on each suggestion would automatically be in line with his other plans and values—and therefore forwent the critical assessment for consistency, leading him into the state of regret that he ended up in.

If we suppose that this variation of case E indeed captures what happened, and that Edward was indeed complacent, let us examine how this affected his synchronic and diachronic self-governance. I will address synchronic self-governance first. To begin, it is a safe assumption that Edward was, in principle, capable of self-governance: there appears to be sufficient unity and organization of the motives of his actions for agential direction, and it seems reasonable to assume that Edward's psychic economy would include self-governing policies (e.g., to give weight to considerations pertaining to his health) for which there were no tendencies or inclinations to change. Therefore, at the moment the first suggestion to stay in on Friday night was made, if Edward believed that staying in would be a means to his end that was consistent with his other plans and his values, and then intended to stay home in part because of his self-governing policy to give weight to considerations about his health, then Edwards's staying in seems to be guided in the right way for synchronic self-governance.

However, now consider a point in time after $t_{E<V}$, when Edward receives another suggestion to forgo an evening out with friends that he follows more or less blindly, without taking into account his evaluative observations of regret from the previous times he followed this suggestion and without resolving the problem that his newly formed intention to stay in poses in relation to last week's intention to go out with his friends again next Friday, which is now *this* Friday. At this point, Edward's synchronic self-governance is blocked because, like in cases A–D, the ensuing action is not guided in the right way by a relevant practical standpoint. After all, by being complacent he has allowed inconsistencies to arise in his relevant plan states that under-

mine where Edward stands on the matter of how to spend his Friday nights.

If this reasoning holds, then we may suppose that Edward's synchronic self-governance would have been blocked at various later times as well in much the same way. As a result, Edward's diachronic self-governance was in effect also undermined by the complacency that was brought about in connection to Edward's employment of the e-coaching system. Like Dave, Edward was potentially diachronically self-governing in other aspects of his life, but considering the regret Edward felt after having reaching his weight loss goal, it is a plausible conclusion that Edward was not diachronically self-governing with respect to the way he spent his Friday nights.

This concludes our examination of example cases in which e-coaching systems facilitated a complacency in people in relation to aspects of their practical reasoning, which in turn negatively affected their self-governance. As I mentioned in the beginning, these cases are not predictions about what will happen, but I do think they adequately illustrate the kind of issues that may arise with the use of e-coaching systems. Cases A-C gave us an initial sense of what automation-related complacency with relation to one's practical reasoning might look like at a particular moment in time. These same cases also helped demonstrate the difficulties involved in determining, on the one hand, a reasonable standard of vigilance, and on the other hand, when to say that someone who drops below that standard of vigilance can be said to be complacent.

With cases D and E, we gained more insight into the temporal aspect of complacency, as well as into the specific roles that the e-coaching technologies can play in facilitating complacency. In case D, we observed how complacency crept in over time after the e-coaching system had gained its user's trust through its persuasive interactions and helpful advice, which subsequently led to a problematic decrease of the user's vigilance in relation to the system's subsequent suggestions. By becoming complacent in this aspect of his practical reasoning, and by simply going along with the e-coaching system's later suggestions, the user allowed means-end incoherence into his plans, which undermined his self-governance.

Then, in case E, we observed how a user's trust in the system grew as he was making progress towards his goal, and how that progress,

together with the e-coaching system's continuous goal-promoting feedback and ongoing encouragements led him into a narrow state of mind in which he became insufficiently vigilant with regard to assessing the e-coaching system's subsequent suggestions in relation to his overall plans and values as these suggestions changed over time in response to the changing circumstances. By being complacent in this regard, this user allowed inconsistency among his intentions, beliefs and plans, which undermined his self-governance.

Taken together, these fictive cases thus reveal how the characteristics of e-coaching systems—their adaptivity to context over time, dialogical abilities, and persuasive techniques (see Section 2.3)—can play into a bias towards automation that can lead people to overly rely on the accuracy of the system and culpably overestimate the extent to which they themselves have put in the effort necessary to ensure means-end coherence and consistency in their intentions and plans. In other words, they illustrate the plausibility of both the empirical claim that e-coaching systems introduce a heightened risk of complacency and the conceptual claim that such complacency undermines self-governance.

Having defended both these claims in this chapter, I hope to have made the case convincingly that the complacency concern is a concern to take seriously. If this is right, then this has normative implications, both for users of e-coaching system and for designers of e-coaching systems. In the following chapter, I will examine those implications.

Chapter 5

Normative Implications and Recommendations

"With information retrieval, anything over 80% recall and precision is pretty good—not every suggestion has to be perfect, since the user can ignore the bad suggestions."

Peter Norvig

In the foregoing chapters I argued that the development and widespread adoption of e-coaching systems raises a number of ethical concerns, including the pressing concern that e-coaching systems could potentially facilitate a self-governance-undermining complacency in people's practical reasoning. In Chapter 4 I elaborated the argument for the complacency concern by drawing on Kawall's work on the concept of complacency and tying in the empirical findings on automation bias and automation-induced complacency. In addition, I explained why complacency undermines self-governance, and used Bratman's planning theory of agency to bring out different ways in which complacency can manifest itself. If my reasoning is sound, then it follows that the complacency concern should be taken seriously by all parties who are aiming to design ethically sound e-coaching systems.

In the present chapter, my aim is to discuss the normative implications that follow from taking the complacency concern seriously. The chapter is split up into two parts. In the first part, I will discuss the normative implications for *users* of e-coaching systems, for *designers*

of e-coaching systems, and for *policymakers* who are responsible for designing and implementing regulations surrounding the use of e-coaching systems. In the second part, I will synthesize the various theoretical considerations about complacency into five concrete recommendations for the responsible design of e-coaching systems, namely to ensure ongoing consent, reveal the reasoning behind suggestions, increase user awareness about system fallibility, offer reassessment opportunities, and promote reflection on suggestions.

5.1 Normative Implications

In the Introduction, I made explicit my assumption that personal autonomy, understood in terms of self-governance, is intrinsically valuable and therefore a normatively relevant concept. Given this assumption, and supposing that the introduction of e-coaching systems indeed introduces a risk of people compromising their self-governance through complacency, it follows that, from a normative standpoint, there is a responsibility to look for ways of mitigating this risk. Notice that this responsibility does not entail a more general commitment to "self-governance perfectionism," in the sense of pushing individuals towards "maximizing" their self-governance. Rather, I take the responsibility to pertain only to a commitment to neutralize the threat to self-governance that is posed by the introduction of e-coaching systems.

With regard to this responsibility, however, it is not unequivocally clear who has to shoulder it. As became evident at the end of the previous chapter, complacency is not something that is attributable in a binary way to either the e-coaching system's users (for letting their guard down) or to the designers (for playing into known decision biases). Rather, complacency comes about through a dynamic interplay between a user's attitudes and dispositions and an e-coaching system's features against a background of that user's socio-economic and political circumstances. As such, it seems reasonable that the responsibility for avoiding complacency should be shared between both parties. After all, placing that responsibility solely with the users might lead to a mismatch between the level of autonomy that the technology requires from people and the actual level of autonomy that they have—something Anderson (2009) has referred to as an "autonomy gap". On

the other hand, placing all of the responsibility with e-coaching system designers would be taking a patronizing attitude towards the users, as it implies a certain helplessness on their part. I therefore propose to take a balanced position that acknowledges that there is an important role for designers to build in mechanisms aimed at minimizing the risk of facilitating complacency, while recognizing at the same time that neither the presence nor the absence of such mechanisms relieves users of their agential responsibility to adopt a critical attitude towards themselves as well as towards the suggestions that are provided externally by the e-coaching system. As I will propose, some of the mechanisms that designers could implement could be geared towards promoting such a critical, reflective attitude. Finally, as e-coaching systems are developed and used within a societal context, there are also implications for governmental agencies that regulate or give counsel about the use of e-coaching technologies. In what follows, I will discuss the normative implications that I see for each of these groups.

5.1.1 Implications for Agents

For agents, the main implication that follows from the previous chapter is that their commitment to self-governance entails that they have to play an active role in ensuring that the e-coaching system to which they have outsourced certain processes is truly working on their behalf. As discussed, this can entail more or less work, depending on the "degrees of freedom" that the system has been given. If agents happen to know the set of recommendations that the e-coaching system can make, then they only have to assess whether those recommendations are fitting for them at this point in time. When an e-coaching system is only reminding an agent of his or her own prior plans, the agent does not have to do a full assessment of the appropriateness of the suggestion (i.e. whether the suggestion fits with his or her values), as the agent has already endorsed the suggestion at some point in the past. Rather, the agent only has to check whether the suggestion fits appropriately with his or her current plans (since those plans may have been updated).

However, the broader and more open the tasks are that an e-coaching system is assigned, the less likely it will be that agents will know what the e-coaching system will recommend. After all, as discussed in

Chapter 2, their level of independence is really what sets e-coaching systems apart from other types of self-regulation facilitators. Therefore, when agents find themselves dealing with external suggestions that they have neither come up with themselves nor seen before, they ought to consider each suggestion as *input* to their practical reasoning, rather than seeing it as an *outcome*, no matter how often the system has steered them right in the past. They have a responsibility to check suggestions for action for means-end coherence, not just in the abstract, but in relation to their own situation. Moreover, they have to assess, at least in a broad way, that the suggestion is consistent with other plans or values they may have.

Let me emphasize once more that this does not mean that one has to engage in deep reflection on one's life each time a suggestion is made. Taking time for those kinds of reflections is important too, of course, and we also have to make sure that the routines that we establish with our e-coaching systems do not mask certain real warning signs that something is wrong in our lives (see also Kjaersgaard (2015); Anderson and Kamphorst (2015)). But that is an altogether different matter. The matter at hand here concerns the idea that outside suggestions for action should undergo at least the same level of scrutiny that our internally generated plans receive by the filtering function of our beliefs and plan-like states. Of course, this will involve *some* effort, but provided that the efforts we save by having the e-coaching system do much of the heavy lifting when it comes to generating options and keeping track of plans, the effort to scrutinize the suggestions will be marginal in comparison.

Finally, it is worth mentioning that these norms are not new; they also govern outside suggestions that we receive from other people. However, the emergence of e-coaching systems has given us cause to make these norms explicit, as the sheer number of suggestions that we can be expected to receive and that we will have to process will increase dramatically if we employ e-coaching systems in various domains of our lives. Moreover, the better these tailoring techniques become, the more we should guard ourselves against being lulled into a false sense of security. Therefore, it is good to draw attention to the responsibilities that agents have towards themselves in regard to being appropriately vigilant. I will say more about the practical aspects of this in Section 5.2. For now, let us turn to the normative implications

that our findings have for designers of e-coaching systems.

5.1.2 Implications for E-Coaching System Designers

As mentioned, finding ways of mitigating the risk of complacency is a shared responsibility. For designers of e-coaching systems this responsibility can be captured in terms of having to resist the temptation to treat tight-knit integration between agent and system as the be-all and end-all purpose of e-coaching system design. That is, contra an attitude often encountered in the literatures on nudging and persuasive systems, the pursuit of the kind of seamless integration discussed in Section 2.4 about extendedness should be curtailed by considerations regarding ways to thwart the risk of complacency. Doing so entails facing two types of practical challenges, both of which concern having to be sensitive to individual's dispositions and susceptibilities.

The first challenge is that designers should be mindful of individuals' biases towards automation—recall the evidence from Section 4.2.1—and try either to reduce the number of complacency-facilitating features, or to add mechanisms that aim to counterbalance the effects of such features. Of course, not all effects can be foreseen, but being aware of the possibility of these types of effects and looking for them specifically in extensive pilot studies is likely to reduce the overall risk of people becoming complacent. In addition, efforts could be made to increase the accuracy of people's attribution of system reliability. In the second part of this chapter I will propose practical strategies for doing so.

The second challenge for designers is to find ways of helping users to shore up their vigilance. Besides implementing a well-thought-out strategy for obtaining informed consent—I will address the subject of "notice and consent" in Section 5.2—I see two avenues that are promising with regard to shoring up people's vigilance through interaction design. The key to both is the observation that users ought to have an active role in determining their individual actions and the course of their lives more broadly. First, special attention should be paid to the phrasing of the suggestions, so as to increase the likelihood of active reflection. For example, instead of suggesting "stay home on Friday night", the system could have explained to Edward (from Section 4.4.3) that his going to bed late on Friday nights was likely interfering

with his weight loss goal and asked him whether he would consider going home earlier (and if so, perhaps a negotiation could happen about how much earlier, cf. Beun, Ahn, Griffioen Both, Fitrianie, and Lancee (2014)). In that scenario, there would thus have been a stronger emphasis on the *dialogical* aspect of e-coaching.

What I take the previous example to illustrate is that the commitment to the importance of self-governance that I am assuming entails an endorsement of dialogue between e-coaching systems and their users, the reasons for which go beyond considerations concerned with effectiveness for behavior change, improved usability, or accurate modeling of human coaches. For the reflections from the previous chapter suggest that, in order to thwart the risk of e-coaching systems facilitating complacency that undermines self-governance, we ought to invest in techniques for dialogue that go beyond the minimal sense of offering choices to which users can respond. Instead, I propose that there needs to be a genuine back and forth between the user and the e-coaching system—something that is sometimes referred to as a *collaborative conversation* (e.g., Anderson and Swim, 1995)—in order to make the user an active participant in the decision making and the plan-making processes. Moreover, the dialogue itself should be directed not just at choosing between suggestions (e.g., would Dave (from Section 4.4.2) want to meet up with a friend for drinks or set up a friendly tennis match?), but at critically assessing suggestions in the first place (e.g., would Dave think that adding social activities to his agenda is right for him?). In other words, the dialogue should be aimed at promoting reflection.

Admittedly, this approach will sometimes undermine some of the intended effectiveness of a behavior change intervention as it also allows people to give themselves significant leeway. And though we typically consider the choice to give oneself leeway or not to be within the decision-making scope of a self-governing agent, we have reasons to think that under certain circumstances such decisions in favor of leeway will be indicative of self-regulation failure (see Section 1.2), especially when decision fatigue is involved or when an agent is in a close-minded, hot state. In those kinds of circumstances it might still be appropriate for an e-coaching system to remind the user of his or her prior plan with the aim of triggering reconsideration of a shifted judgment, but it might be counter-productive to ask the user

to engage in effortful deliberations about new plans as he or she will be more likely to choose short-term gains over long-term benefits.

One way of putting this is that it would be prudent if design specifications of e-coaching systems included the aim of prompting people to deliberate at times when they are not obviously pre-engaged. I formulate it in this way, since, realistically speaking, it will prove to be exceedingly difficult, if not outright impossible, to accurately assess when someone is "really" in a good (or "the best") position to engage in deliberation. By framing it negatively, the aim simply gives expression to the idea that one should not be prompted for deliberation at times when one is obviously pre-engaged or already in the situation that the prompt is about, to the extent that the system can know about this. For example, in Edward's case, it would mean not prompting him about whether or not to stay home when he is on the phone with his boss or when he is already at the bar with his friends.

The second way in which e-coaching systems could be designed to promote critical engagement is by following up on suggestions with questions *about the suggestion* that prompt people explicitly to reflect on them. This allows agents to explicitly reject or endorse certain suggestions or types of suggestions. In Dave's case, one or two evaluative questions at appropriate moments about his experience of going out with friends and playing tennis would have provided him with opportunities to realize in a much earlier stage that these activities were not benefiting him. Moreover, he would then have been in a position to inform the system that he did not want any more suggestions in that direction.

As mentioned, in the second part of the chapter I will offer a number of recommendations for system designers based on the normative implications discussed here. As a conclusion to this first part, however, let us turn our attention to how the complacency concern affects how policymakers and regulators should approach the decision making about the use of e-coaching systems.

5.1.3 Implications for Policymakers

As we have seen in Chapter 1, there is currently a trend towards the use of e-coaching systems in a variety of domains, driven by hopes of tackling both individual and societal problems. Within this trend,

there are proponents who argue for the use of incentive programs for e-coaching systems, or even for instantiating mandatory use in certain circumstances. As a result, policymakers and regulators who are concerned with issues such as health and well-being, but also sustainability, and public safety, will have no choice but to face difficult decisions about the use of e-coaching systems. The pertinent question is how they should approach these decisions.

One normative implication of what I have argued for in this dissertation is that policymakers and regulators should apply a cautious approach in relation to incentive programs as well as to proposals for mandatory use of e-coaching systems. Given the value of self-governance, such a principle is warranted, I believe, in light of the many unknowns that there currently are about the ways in which e-coaching systems can potentially undermine people's self-governance, including the ways I have suggested in the previous chapter about complacency.

Recommending a cautious approach with regard to policies about promoting the use of e-coaching systems is to leave open the possibility for exceptions, and this is by design. For the claim is neither that there is no place for e-coaching systems at all, nor that e-coaching systems cannot serve specific purposes. In fact, I suspect there may be particular domains in which mandatory use of e-coaching systems could be justified. For example, it might be appropriate to require the use of e-coaching systems for patients whose personal autonomy is already compromised (e.g., people who suffer from early symptoms of dementia), with the intention not to reduce the level of care they receive, but to allow them to live more independently (for a good example, see Mihailidis, Boger, Craig, and Hoey (2008)). Another group where mandatory use of e-coaching systems could potentially be appropriate, especially if they implement "anti-complacency mechanisms" (see Section 5.2), is with inmates or convicted criminals on parole. In many correctional facilities, inmates already receive coaching or mentoring, and people on parole have to keep in touch with their parole officers, and are often also required to attend counseling sessions. E-coaching technologies could form a good addition to already existing measures aimed at relapse prevention. As such, it could potentially be argued that in such contexts the mandatory use of e-coaching systems could be justified, provided it was deemed a suit-

able, necessary, and reasonable measure (following the proportionality principle (Harbo, 2010)).

These are vexed matters though that deserve careful treatment, and the decisions about these matters are much more complex than I am portraying them here. They do serve my illustrative purposes, however, in showing that taking a cautious approach is not to flat out deny that there may be legitimate reasons for regulators to proscribe the use of e-coaching systems. Rather, the point of a cautious approach in the context of discussions about e-coaching systems is to underline that, if people consent, it should be on the basis of their own active decision to share data and to have one's behavior be influenced by an e-coaching system. By placing the burden of proof with those who wish to promote the use of e-coaching systems in a particular domain, the aim is to protect individuals' self-governance and liberty (see Section 3.1) and to avoid that policymakers and regulators become fixated on self-regulation support systems such as e-coaching systems as automatic, catch-all solutions to a plethora of issues that may require very different solutions.

Finally, with regard to policy-making, there is the further matter to consider of the current proliferation of self-help and self-management tools that are being brought into various domains by consumers. In relation to health, for example, people are looking for applications that can assist with improving their (mental) health without seeing a specialist (e.g., apps that provide stress reduction exercises). In contrast, others want their practitioners involved and for them to have direct access to the behavioral patterns that their apps record in the hope of receiving care that is better tailored to their specific situation. Health care professionals on the other hand can be wary when it comes to integrating data streams from consumer products into their daily practice, as they will typically have very little evidence of the validity of these data.

These developments are changing the ways in which patients and caregivers interact, and the expectations that both sides have of one another. Because of this, I foresee a call for regulation, so that the new way of interacting between patients and health care professionals can be given shape and mutual expectations can be streamlined.

A major challenge here is that the benefits and limitations of using e-coaching systems differ tremendously depending on the features

of a particular system. With the proliferation of apps, policymakers and regulators will have a difficult time deciding which apps to endorse. Given this challenge, I think it will be crucial to establish an independent review institute that has the expertise to hold e-coaching systems against certain standards—including a measure of how complacency facilitating a system is—and can give out "quality certifications". Within the Netherlands, several of such quality insurance programs have been instantiated for eHealth products more generally, such as TNO's "Ehealth analyse en sturingsinstrument" (eASI) (Mikolajczak, Blanson Henkemans, and Keijsers, 2011) and the "Onlinehulpstempel" from the Dutch Trimbos Institute. While commendable, these initiatives focus predominantly on establishing a trustworthiness score of self-help applications based on an estimate of effectiveness and usability. My proposal, on the other hand, would be to also have an institute—perhaps with a supervisory board whose members are university-employed academics (computer scientists, ethicists, psychologists)—that authorizes certificates or labels to systems that take adequate measures to counteract automation bias, and promote critical reflection and vigilance, thereby minimizing the risk of complacency.

Some might worry that this proposed measure is not far-reaching enough with regard to protecting individuals from influences that may undermine their self-governance. Instead, they might argue that the complacency concern calls for the founding of regulatory agencies that protect people from themselves by determining which products can be brought onto the market, comparable to how the U.S. Food and Drug Administration regulates which drugs may be legally sold in United States. After all, establishing a review institute by itself does not alter the possibility of complacency-facilitating e-coaching systems being introduced. In fact, developers and consumers would be able to bypass "anti-complacency mechanisms" altogether should they choose to, for example if developers would find those mechanisms difficult to implement, or if consumers would find the mechanisms tedious and interfering with usability. Regulatory agencies, on the other hand, could enforce a ban on "unlicensed" e-coaching systems.

With the current state-of-the-art of e-coaching systems in mind, I think this position is currently unwarranted by the available evidence. For while the uncertainties regarding the use of e-coaching systems

and the negative effects this use may potentially have for people's self-governance warrants a cautious approach from policymakers with regard to proscribing these systems, they do not, at present, seem to warrant a limitation on people's liberty to choose to employ certain kinds of systems with the aim of promoting their own welfare (see again Section 1.3). It is difficult to foresee whether a different approach will be necessary for future generations who grow up using e-coaching systems and whose identities may therefore be linked much tighter to these technologies, but as it stands, having a review institute in place to provide clear information about the quality of self-regulation support systems will already go a long way in helping people to assess for themselves which technologies they want to use, and how much of the shared responsibility to fend off complacency they want to shoulder. Moreover, and equally importantly, a review institute will help policymakers and regulators to make informed decisions in contexts in which strict regulations about the use of e-coaching systems will be warranted.

This brings a close to the normative implications for policymakers and regulators. Much more research is certainly needed to assess the size of the complacency risk and to learn more about the ways in which complacency can be combatted and prevented. Throughout the next section I will highlight a number of specific research questions that, taken together, can be viewed as an initial research agenda. Until progress is made in this regard though and more information becomes available, I think it is best to err on the side of caution when it comes to policy-making surrounding the use of e-coaching systems.

With the normative implications now on the table, it is time to become more concrete. In the remainder of the chapter I will use the normative ideas that were discussed to make five recommendations for designers and developers for combatting the facilitation of complacency.

5.2 Anti-Complacency Recommendations

The recommendations that I will make here aim to be both general enough to have wide application and concrete enough to serve as guidelines for designers and engineers. While considering these recommendations, it will be important to keep in mind that they are

conditional on my arguments about the complacency concern holding water but are made independently from the various other potential concerns as discussed in Chapter 3, as those require further elaboration and careful consideration beyond the scope of this dissertation. The following is therefore restricted to a set of "anti-complacency" recommendations, each of which will be introduced with a brief motivation and followed by some elaboration.

Obtaining Ongoing Informed Consent

Let us begin by reflecting on some considerations regarding informed consent to being subjected to certain influences (Faden and Beauchamp, 1986), since questions about consent will likely mark the first interaction people will have with e-coaching systems. As such, these interactions will likely be the first opportunity to create awareness at the outset about the system's use of persuasive techniques and the probabilistic and therefore fallible nature of the system's recommendations, which could already be a strong move in the direction of shoring up people's vigilance with regard to assessing those recommendations (cf. Bahner et al. (2008)). Moreover, obtaining informed consent is also more generally important in relation to respecting people's personal autonomy by providing users with a genuine opportunity to give or deny their approval to being subjected to certain influences, be it medical interventions or behavioral recommendations.

Unfortunately, informed consent is a notoriously tricky aspect of product deployment to get right: on the one hand, these documents often end up including technical jargon or "legalese" that severely hinders comprehension (e.g., Stunkel, Benson, McLellan, Sinaii, Bedarida, Emanuel, and Grady, 2010), and, on the other hand, people often do not seem to understand the relevance of these documents (cf. Faden and Beauchamp (1986, pp. 300–302) on the mistaken presumption that people understand informed consent as an act of authorization). For example, even with regard to something as crucially important as experimental cancer treatment, one study (N=287) found that 70% of patients did not understand the unproven nature of the treatment, and 63% of the patients did not understand the potential for incremental risk from participation (Joffe, Cook, Cleary, Clark, and Weeks, 2001).

Such findings highlight the difficulty of obtaining actual informed

consent instead of a meaningless signature or, in today's digital age, the click of a button. Another problematic aspect is that obtaining informed consent is traditionally viewed as a one-time occurrence: once consent is given it is assumed to last forever, or until it is rescinded. In any case, it is rarely renewed (Custers, 2016). In the case of e-coaching systems, this is particularly troublesome as these systems are designed to be adaptive and to use different persuasive techniques and behavior-guiding strategies over time. In addition, new ways of using the user's already collected data may be introduced at a later time when an update of the e-coaching system is installed (cf. Barocas and Nissenbaum (2014, p. 59–61)). As such, it may not be sufficient to inform users at the very beginning of how they operate, as this will likely be either on a level of explanation that is too abstract or too detailed (the *transparency paradox*, see Nissenbaum (2011)). What would be better, I believe, is if e-coaching systems would strive to secure users' *ongoing consent*. After all, it will be exceedingly difficult for users to foresee how they will experience the interactions with their e-coaching system given that the e-coaching system adapts over time. The first recommendation then, is the following.

> **Recommendation 1: Ensure Ongoing Consent.** *If e-coaching systems are to counteract automation-related complacency, they should, before actual use, inform their users about the use of persuasive techniques to influence the users' behaviors, test the user's comprehension of the information provided about these techniques, and secure consent for using these techniques. Moreover, e-coaching systems should, during use, routinely secure the consent from their users to continue using their persuasive techniques.*

Implementing the first half of the recommendation is a challenging task in itself, but one for which it is possible to draw on existing practices for developing more "traditional" informed consent procedures. Relevant questions here are what the appropriate standard of awareness should be, what information should be divulged to that end, and how to assess whether the user has sufficiently comprehended the information provided (cf. Faden and Beauchamp (1986, pp. 327–328)). In this context too a "relevant person standard" could potentially be utilized to determine the standard of awareness people should meet before they can go ahead and use the e-coaching system, as well as for

determining the information that users would want to have in order to make an informed decision (see, e.g., Greenblum and Hubbard, 2019).

With regard to the ways in which the information should be presented and the user's comprehension tested, it could be sufficient, especially in non-clinical domains, to implement a mandatory "first start wizard" which guides people through a number of key principles and techniques, explains the risks involved with using e-coaching systems (including the risk of becoming complacent), and checks comprehension through a series of well-constructed key questions. While this kind of "recognition test" will have its limitations, it may nevertheless be useful as a screening tool to identify people who may be of need of further support in relation to using the e-coaching system in question.

In health care domains on the other hand, it might be more appropriate to involve a health care professional in going through these steps, to increase the chance that users truly understand to what they are consenting. Here, a form of "feedback testing" could be used by the health care professional (Faden and Beauchamp, 1986, p. 328), for example by asking users to restate in their own words what has been disclosed to them. In any case, users should be made aware of the independence of e-coaching systems, so that they realize that suggestions, even though they may be personalized, can be inappropriate for them at particular times.

Realizing the second half of the recommendation, the part about *ongoing* consent, will require some more ingenuity. That is because simply asking for consent at a routine interval (e.g., using expiry dates (Custers, 2016)) will likely lead to annoyance with the system, which might lead to a decrease in usability, and, ultimately, effectiveness. One might be tempted to think that incurring this cost is the price to pay for ensuring active reflection by the agent on the influences that the system might have on his or her behavior and cognitive state. The problem, however, is that it is not evident that repeating the request for consent will actually have this effect. What is more likely to happen, I believe, is that people will automatically recognize the similarity of the question and answer it mindlessly out of habit.[1] This, of course, is not what the prompt is meant to accomplish.

Rather, the desired outcome is for this particular type of request to trigger conscious thought at a time that is both opportune for the

[1] See also the literature on habituation (e.g., Thompson, 2009).

user and also meaningful with respect to the coaching process. Thus, given that the system already learns new patterns over time and may employ different coaching strategies depending on what it is has learnt, it seems opportune to view those strategy changes as opportunities for meaningful reevaluation of consent since different strategies might well involve other sources of data and different persuasive techniques.[2]

An open research question here is whether it would be best to prompt at the point just before the switching is about to happen (which we might call prospective prompting) or at a point in time a brief period after the system has switched (evaluative prompting). The benefit of prospective switching is that people are given the opportunity to put a halt to a certain type of influencing before it occurs. The downside, however, is that people will have to make this decision without having experienced this type of coaching. The alternative is to rely on the initial, blanket consent for exploring a new coaching strategy, but to ensure ongoing consent after a brief period of trying this new strategy with evaluative prompting.

As mentioned above, further research and reflection is needed to determine which of these ways of attaining ongoing consent is preferable. The benefit of either of these ways as opposed to the time-based prompting approach is that there is a natural opportunity for making the consent prompts informative and that it allows the consent question to be integrated in the ongoing dialogue between the user and the e-coaching system. The hope is that this will provide an intuitive interaction, without making it easy for users to give consent mindlessly.

Having addressed this more general issue of consent, let us now turn our attention to more specific ways of counteracting the complacency concern. As we have previously seen in Section 5.1.2, system designers can combat complacency in two ways. On the one hand, they can take measures to minimize the effects of automation bias. On the other hand, they may add features which facilitate vigilance in their users. After all, the fact that users have an agential responsibility to remain appropriately vigilant with regard to the system's suggestions

[2]Note that the notion of a strategy change should not be equated with a subtle change *within* a particular strategy. Learning that someone likes to receive his notifications fifteen minutes earlier than other people in similar situations and adapting to this preference is not a strategy change, whereas changing the coaching content in relation to a previously unexplored factor would be.

does not mean that they cannot or should not be supported in these efforts. I will discuss recommendations relating to each of these design challenges in order.

Combatting automation bias

As mentioned in the first part of the chapter, different approaches can be taken to counteract the risk of facilitating complacency (for overviews, see Alberdi et al., 2009; Goddard et al., 2012). Not all of these approaches are directly applicable to e-coaching systems however. Some approaches, such as changing the location of warning signals on screens are very specific to contexts in which users continuously interact with decision support systems to perform certain highly demanding tasks (e.g., controlling air traffic). Other approaches, such as letting users have a negative experience with the system in order to reduce trust in the automation and thereby reducing the effects of automation bias (e.g., Bahner et al., 2008) are quite evidently problematic when it comes to coaching, especially in health domains.

Fortunately, there are also two approaches that stand out positively because they fit well with the nature of coaching. As we have seen in Chapter 2, coaching is not about telling a coachee what to do, but rather about having a collaborative dialogue in which a coach assists a coachee in forming appropriate goals and crafting realistic action plans. This process therefore naturally involves *explanation*: if coachees are to further develop their self-regulation skills, they ought to understand why a coach makes the suggestions that he or she makes, and this involves gaining insight into the reasoning processes—including reasoning about assumptions and uncertainties—that lead coaches to certain suggestions.

In light of this, the first strategy I find compelling is making users aware of the reasoning process behind a suggestion in order to reduce the chance that they will blindly follow the system's advice because they overestimate the system's expertise (Goddard et al., 2012, p. 125). If users can inspect the reasoning behind certain suggestions, then they at least have the opportunity to reflect on this reasoning. As an additional benefit, previous research in the domain of human-agent teamwork has shown that the ability to provide such insights improves user experience (e.g., Harbers, Bradshaw, Johnson, Feltovich, Van den

Bosch, and Meyer, 2012). The recommendation, then, is this.

> **Recommendation 2: Reveal the Reasoning Behind Suggestions.** *If e-coaching systems are to counteract automation-related complacency, they should be able to provide insight into the reasoning behind their suggestions.*

The word "insight" here is meant to indicate both that the explanation behind a suggestion should be non-trivial—i.e. that it goes beyond merely stating that the suggestion is made in light of the user's goals or something to similar effect—and that the explanation can be grasped by laypeople. While striking a reasonable balance between these two aims may be straightforward for human coaches, it actually poses a significant research challenge with regard to e-coaching systems. This is because, historically, architectures for agent systems in which explanation of reasoning is a core component—e.g., Bratman, Israel, and Pollack (1988); Rao and Georgeff (1995); Broersen, Dastani, and Van der Torre (2005); Harbers, Van den Bosch, and Meyer (2010)—have been criticized for lacking the ability for learning and being adaptive.

On the other hand, systems that rely solely on machine learning techniques for making suggestions will have difficulty implementing this recommendation, as many of today's machine learning techniques do not, out of the box, provide human-readable reasoning steps. Typically, the class of algorithms used for learning and predictive modeling uses architectures that involve a set of numeric weights that are adapted to best fit a set of training data. As these numeric weights resist straightforward interpretation, it is often difficult to describe, even for people who are well versed in the workings of these types of algorithms, exactly how a particular prediction came about.

The upshot of all of this is that, in order for e-coaching systems to provide insight into their reasoning, while also harnessing the power of machine learning to be adaptive, research efforts will be required to integrate these two approaches. Significant efforts are being undertaken under the heading of "Explainable A.I." (XAI) to do just that (e.g., Biran and Cotton, 2017; Abdul, Vermeulen, Wang, Lim, and Kankanhalli, 2018; Hall and Gill, 2018; Meacham, Isaac, Nauck, and Virginas, 2019). Developers and designers of e-coaching systems are therefore advised to track these developments closely.

Before continuing to the next recommendation, there are two final remarks I want to make about this one. The first is that implementing

the recommendation does not strictly require that *every* suggestion needs a precursory statement about the reasoning involved. While the information should be available, it is defendable that in many cases the information would remain hidden unless it is requested by the user. That way, the focus of the ongoing dialogue can remain with the message that the system is trying to convey. That said, however, the information should also not be hidden by default, as that would undermine the attempt to combat complacency. A good solution I think would be to show the reasoning behind a new suggestion (one that the user has not seen before), and to prompt the user when the reasoning behind a suggestion changes (even if the suggestion itself has previously been endorsed by the user).

The second remark is that the interface in which the reasoning is shown could be connected to a mechanism that allows users to identify flaws in the reasoning and to quickly provide feedback to both the system and the software developers. That way, users partake actively in shaping the coaching process.

Besides providing insight into the system's reasoning process, the second promising approach to minimizing automation bias that fits well with the nature of coaching is to give users insight into the level of confidence that a system has in a particular suggestion. The idea here is that this should reduce unjustified compliance as it allows users to determine for themselves which confidence levels they find acceptable. Moreover, the mere fact that these confidence levels can fluctuate would make users aware of the fallibility of the system. The recommendation then, is this.

> **Recommendation 3: Increase User Awareness of System Fallibility.** *If e-coaching systems are to counteract automation-related complacency, they should make users aware that suggestions are made on the basis of assumptions and probabilities that may be inaccurate, and that a system's suggestions can therefore be more or less appropriate to the user's current situation.*

Wickens et al. (2015) suggest that one good way of doing this would be to display a "confidence score" with each suggestion, which would represent the level of confidence yielded by the e-coaching system's algorithms that the suggestion is fitting for the user. This approach has yielded promising initial results with decision support systems

in a study with pilots, where people in the experimental group were supported by a support system which displayed a confidence rating on their screens. The results from this study show that this minor change in the interface was related to less automation-induced mistakes by the pilots in that group (McGuirl and Sarter, 2003). Similarly, then, it may be hypothesized that showing a confidence score with each suggestion that an e-coaching system makes is likely to make users less susceptible to becoming complacent by reminding them of the system's fallibility and thereby subtly priming them to consider the suggestions on a case-by-case basis.

Despite the promise of this approach, there are a number of important research questions that will need to be examined carefully. One such question pertains to the way in which people will respond to appropriate suggestions that are accompanied by low confidence scores. This may for instance occur in the early stages of an interaction with an e-coaching system when the system has not yet learnt the user's specific preferences, but bases its suggestion on some form of aggregated population data. One potential outcome is that people will be hesitant to follow these suggestions despite their appropriateness. By itself, this hesitancy may be welcomed as a form of vigilance, but if the result of low confidence scores is that people are discouraged to interact further with the e-coaching system, then this would likely hinder the success of the coaching process.

Another reason to thoroughly test confidence scores empirically is to check for the potential issue that consistently high confidence scores could actually facilitate excessive trust in a system. Though the results from McGuirl and Sarter suggest that users are more likely to examine each suggestion on a case-by-case basis, it is not evident that this behavior will remain stable as time goes by and users become accustomed to high confidence scores. In light of this potential issue, it could prove to be worthwhile for systems to be transparent about the way in which confidence scores are generated, so that users do not have to take it on faith that the scores are accurate, but that they instead have an opportunity to assess the credibility of the confidence scores for themselves. How to best implement such transparency is an open research question in itself in need of careful study (see again the point about Explainable A.I.).

Clearly, the recommendations for combatting (the effects of) auto-

mation bias that I have discussed in this section do not exhaust all potential strategies; other possibilities for effective countermeasures may be developed as the body of work into automation bias grows. However, I do think that for e-coaching systems these will be a good starting point. Let us now turn to the second challenge that system designers face, namely to find ways of supporting people in maintaining, or even shoring up their vigilance.

Supporting users' vigilance

The primary way that I see of supporting users' vigilance is by implementing mechanisms in e-coaching systems aimed at actively stimulating critical reflection by their users. As these mechanisms can take different forms, and stimulating reflection can be done at different times and in different ways, this statement deserves elaboration.

To begin, recall that part of what it means to be a self-governing agent is that one is in charge of mapping out a course for one's own life. This process involves reflecting on what it is that one truly cares about. Part of the complacency concern is that the smooth and fluid user experience that future e-coaching systems are likely to offer, might run counter to this important aspect of self-governing agency. After all, the thought is, if there are fewer moments in which people experience dissonance, there will be fewer triggers that will give cause for serious reflection. To counter this concern, designers of e-coaching systems would do well to implement ways of offering genuine opportunities to their users to reassess whether they actually want to be coached, and if so, in what direction. The corresponding recommendation is the following.

> **Recommendation 4: Offer Reassessment Opportunities.** *If e-coaching systems are to counteract automation-related complacency, they should offer users genuine opportunities for reassessing their commitment to goals that the e-coaching system supports, as well as to using the e-coaching system in general.*

Importantly, when offering users these opportunities for reassessing their commitment, the e-coaching system should be as neutral and open as possible in the phrasing (as opposed to being persuasive). In addition, some of what was said about the consent recommendation

also finds applicability here. That is, implementing this recommendation will require efforts to ensure that people do not habitually (and possibly complacently) answer "yes" to prompts asking if they still want to continue with a particular coaching program. Rather, it is important that these particular prompts will be designed in such a way that that they trigger people into a reflective mode of agency to answer the questions that are being asked of them.

But while we do not want people to answer "yes" out of habit, we also do not want users to answer "no" out of frustration. This means that, just as with the consent questions, it will be crucial to get the timing right. People will need to be in an appropriate environment as well as in an appropriate state of mind to engage in this type of thinking. As mentioned in Section 5.1.2, a starting point here could be to avoid prompting for deliberative tasks at times when one is obviously pre-engaged or already in the situation that the prompt is about. That said, herein clearly lies a major research challenge, one that has been receiving more and more attention recently (e.g., see Aminikhanghahi, Fallahzadeh, Sawyer, Cook, and Holder, 2017; Bidargaddi, Almirall, Murphy, Nahum-Shani, Kovalcik, Pituch, Maaieh, and Strecher, 2018), but still requires further investigation.

In practice, questions relating to overall commitment as well as to goal commitment could perhaps be integrated in the same conversation that addresses ongoing consent (see again Recommendation 1). Supposing at least that a good moment has been identified, it might be opportune to also openly address the bigger question of whether they want continue at all. However, conceptually, it is good to separate the question of whether someone is willing to be subjected to certain treatment, and whether someone is still genuinely committed to a particular goal or coaching program.

Lastly, in line with but distinct from the previous recommendation, e-coaching systems should incorporate mechanisms to promote critical reflection not just about the goals and coaching in general, but also about the suggestions that are being made. More specifically, the recommendation is this.

> **Recommendation 5: Promote Reflection on Suggestions.** *If e-coaching systems are to counteract automation-related complacency, they should promote critical reflection concerning the suggestions that the system proposes, particularly about whether*

> *suggestions are fitting for the individual with respect to the goals the individual is pursuing as well as in relation to the individual's values.*

Notice, first, that this recommendation is not a blanket endorsement of ever more deliberation. As Bratman (see Section 4.1) continuously stresses, ever more deliberation is often impractical or even outright impossible given the type of beings that we are (1987, see also Harman (1986)). Rather, the recommendation is about getting users to engage in sufficient deliberation to ensure that they are following a suggestion because they consider it a fitting suggestion for them in their situation, and not simply because the system tells them to.

The insight driving this recommendation is the one we gained from the cases of Dave and Edward from Sections 4.4.2 and 4.4.3. In Dave's case, we saw how the system made recommendations that were seemingly and plausibly in pursuit of the end of stress reduction—they may well have benefitted another individual working towards a similar goal—but that were nevertheless not promoting means for Dave to reach his end. And, in Edward's case, we saw how means-end coherent suggestions (helping Edward to lose weight) could still not be fitting for an individual if the suggestions are not consistent with other plans and values that the individual may have (spending time with his friends).

The lesson from these two cases is that e-coaching systems will operate, by necessity, within a limited scope, and will therefore have significant blind spots when it comes to the complex inner workings of the individual. Taking this seriously means explicitly acknowledging the user as an active participant in the coaching process by inviting him or her to think about, and perhaps share, his or her own unique perspective on the situation.

In addition, as mentioned in Section 5.1.2, the system could at times—though sparingly—ask explicit evaluative questions about suggestions that were new to the user. Doing so would also allow the user to explicitly reject a type of suggestion, which, as it happens, is also preferable with relation to the e-coaching system's learning processes as opposed to dealing with suggestions that were simply ignored (which could have a variety of causes).

Lastly, e-coaching systems could offer some form of "complacency feedback," in which the system responds to certain behaviors that

function as proxies for complacency. As a hypothetical example, if a user were to continuously endorse default options, the system could require a forced choice question every once in a while, explaining the importance of the user's own participation in the coaching process. Of course, it could happen that the proposed default options just happened to be a good fit for this individual, but if the hypothesis holds that agreeing quickly to default options will often be indicative of complacency, then it would be worthwhile nonetheless to take the time every so often to make sure that the user is indeed still actively engaged in the coaching process in such a subtle way.

This kind of complacency feedback approach has not yet been tested, in part because, at present, it is not evident which behaviors could serve as proxies for complacency. As such, there is a call for further reflection on the behavioral aspects of complacency in order to generate hypotheses about proxies that could subsequently be tested empirically. Despite the lack of empirical evidence, I mention the strategy here anyway because it highlights another domain where future research may uncover promising strategies for pushing back against complacency-facilitating automation by helping users develop and maintain an appropriate level of vigilance. And perhaps, if ethical concerns about diminishing self-governing agency are taken seriously in the development of e-coaching technologies, then one day these technologies could turn out to be an essential asset in learning the necessary reflective skills needed for being a self-regulating, self-governing agent.

Conclusion

I began this dissertation by sketching a future in which people are seamlessly supported in their self-regulation efforts by advanced technological systems. The sketched scenarios were hypothetical, but since writing them down, efforts have continued, both in industry and in academia, to further develop technologies in the direction that will make those scenarios, and many others like them, a reality.

My aim with this dissertation has been to contribute to the responsible innovation of self-regulation support technologies, by staging interventions in various relevant debates, sketching the ethical landscape surrounding these technologies, and by offering practical guidelines for various stakeholders for developing, using and regulating these technologies. In this final section, I will recapitulate the main points from each chapter, and conclude with a discussion about the specific contributions I think the dissertation makes.

In **Chapter 1**, I began by distinguishing self-regulation failure from the two philosophically more familiar notions of akrasia and weakness of will, concluding that self-regulation failure captures a wider variety of agentic failings and reveals more of the complexity of the causal mechanisms underlying these failings. I then set forth my views on the concepts of self-regulation and self-regulation failure, arguing that these concepts are more constrained than how they are sometimes portrayed in the literature. Next, I touched upon potential interventions and the role that new technologies can play in offering around-the-clock support that is fitted to individuals' specific preferences and behaviors. I then discussed three perspectives—viz. the perspective of social impact, the perspective of individual health and well-being, and the perspective of self-governing agency—from which arguments can be made in favor of developing technological interventions to bolster people's self-regulation. Finally, I emphasized that these reasons in

favor of self-regulation support technologies are only *pro tanto* reasons that may be outweighed by reasons stemming from ethical concerns that arise with the widespread adoption of these kinds of technologies.

In **Chapter 2**, I defined "e-coaching systems" as a set of computerized components that constitutes an artificial entity that can observe, reason about, learn from and predict a user's behaviors, in context and over time, and that engages proactively in an ongoing collaborative conversation with the user in order to aid planning and promote effective goal striving through the use of persuasive techniques. I argued that focusing on this specific subset of self-regulation support systems brings into view a distinct set of ethical concerns that arise as a result of these systems' level of sophistication and independence. As a conclusion to the chapter, I characterized the socio-technical relationship between user and e-coaching system in light of the literature on the extended mind and the extended will. In anticipation of the argument to come about complacency, I suggested that the kind of tight-knit integration necessary for speaking of an extended system of "user plus e-coaching system" might actually not be desirable for individuals if it entails abandoning certain vigilance-related agential responsibilities.

In **Chapter 3**, I provided an initial inventory of the specific ethical concerns that the widespread adoption of e-coaching systems raise. I distinguished three categories of concerns: concerns about social justice, in particular about equal access, liberty and welfare; concerns about infringements of rights of autonomy, in particular about privacy as well as about coercion and manipulation; and finally concerns about diminishing self-governance as a result of the interplay between e-coaching technologies and their users, in particular the concern that e-coaching systems can potentially have negative effects on people's exercise of self-governance via the facilitation of complacency in people's practical reasoning. The chapter concluded with a first sketch of this complacency concern.

In **Chapter 4**, I introduced Bratman's Planning Theory of Agency as the model of self-governing agency that I would be using for my analysis of the complacency concern. With this conceptual framework in place, I then elaborated on two central claims. First, I introduced the *Heightened Risk of Complacency Claim*, and subsequently highlighted key findings from the empirical literature on automation bias and automation-induced complacency to argue that people likely have a

bias in relation to technology to overestimate the reliability of automated systems. As I observed conceptual confusion in the literature about the concept of complacency and its relationship to automation bias, I subsequently introduced Kawall's more technical definition of complacency and showed how that definition helps to distinguish between having a bias and allowing oneself to go along with this bias and thereby making an epistemically culpable overestimation of how well certain agential responsibilities have been met by relying on the external system, such that this overestimation produces a problematic lack of appropriately motivated, appropriate effort in relation to actually fulfilling those responsibilities. Applied to e-coaching technologies, I thus argued that the risk with e-coaching systems is that they would provide such a smooth, fluid, reliable user experience combined with expert knowledge that they might lull people into a state in which they overestimate how much they themselves have accomplished by employing the e-coaching system with respect to making sure that one's actions are in line with and because of one's own values.

Second, I introduced the *Implication for Self-Governance Claim*, and argued that self-governance involves having a practical standpoint and that the process of forming a practical standpoint requires a certain vigilance and diligence with regard to assessing both internal and external influences. If this holds true, I subsequently argued, then it entails that if an agent is overly lenient, and acts on an e-coaching system's suggestion that she did not assess in light of her constellation of beliefs, intentions and plans, her synchronic self-governance is undermined because her action is not appropriately linked to her practical standpoint. Moreover, by dropping her vigilance with regard to assessing an e-coaching system's suggestions, her diachronic self-governance is also undermined by allowing inconsistencies or means-end incoherences to be introduced into her constellation of beliefs, intentions and plans. To drive the complacency concern home, I concluded the chapter with a series of fictional cases in which I identified different ways in which different functional components of e-coaching systems can contribute to the facilitation of self-governance-undermining complacency in people's practical reasoning.

Finally, in **Chapter 5**, I worked out the implications of the complacency concern for users and designers of e-coaching systems, as well as for policymakers. *With regard to users*, I argued that their commit-

ment to self-governance entails a responsibility to play an active role in ensuring that external suggestions that they have neither come up with themselves nor have seen before undergo at least the same level of scrutiny that their internally generated plans receive. No matter how often the system has steered them right in the past, users ought to check suggestions for action for means-end coherence in relation to their own situation, and ascertain, at least in a broad way, that the suggestions are consistent with other plans or values they may have.

With regard to designers of e-coaching systems, I argued that they ought to be mindful of individuals' biases towards automation by trying to either reduce the number of complacency-facilitating features, or by adding mechanisms that aim to counterbalance the effects of such features. Moreover, I argued that designers should find ways of helping users to shore up their vigilance, mainly by focusing on making the dialogue between system and user a true collaborative conversation in which suggestions for action are balanced with evaluative questions about the suggestions to ensure that the user is appropriately engaged in co-shaping the coaching process.

With regard to policymakers, I underlined the importance of taking a cautious approach in relation to incentive programs as well as to proposals for mandatory use of e-coaching systems. In addition, I argued for establishing an independent review institute that has the expertise to hold e-coaching systems against certain standards—including a measure of how complacency facilitating a system is—and can award "quality certifications". Such a review institute, I argued, will be beneficial for individuals with regard to assessing for themselves which technologies they want to use, and for policymakers and regulators to make informed decisions in contexts in which regulations about the use of e-coaching systems are warranted.

In the final part of the chapter, I made five concrete recommendations for mitigating the complacency concern and highlighted related research questions. **First**, in light of the adaptiveness of e-coaching systems to users' changing circumstances and behaviors, and their potential to change persuasive strategies over time, I suggested going beyond the traditional "notice and consent" strategy and researching ways of ensuring ongoing informed consent by developing and incorporating at appropriate times a coherent bundle of information snippets, comprehension tests and consent questions about the persuasive

techniques relevant to the current state of an individual's coaching process. **Second**, to allow people to better assess the validity and appropriateness of suggestions, I recommended that e-coaching systems would provide insight into the reasoning behind suggestions, thereby adding further support to the upcoming research field of "Explainable A.I.". **Third**, as another measure to counter the tendency to overestimate system reliability, I recommended increasing user awareness about system fallibility. Here, I suggested using a kind of confidence score, but at the same time highlighted the open research questions about such approaches in relation to e-coaching systems. **Fourth**, to support users' vigilance, I recommended implementing explicit reassessment opportunities, both in relation to users' commitment to specific goals and to their commitment to using the e-coaching system in general. The main research question I identified related to this recommendation is how to determine the appropriate moment for prompting a user for this kind of reflection. **Fifth**, and finally, I recommended promoting a form of reflection on suggestions by accompanying a suggestion for action with an invitation to think about whether the suggestion suits the individual user, and by (sparingly) asking evaluative questions about previous suggestions. I emphasized that this recommendation (or any of the others for that matter) is not a plea for ever more deliberation, but rather a way of structuring the user interactions so that they become and remain a form of two-way communication that benefits the coaching process and counteracts complacency in people's practical reasoning.

Having looked back, we are now in a position to see clearly the contributions that this dissertation makes. The first contribution is comprised of the conceptual clarifications the dissertation introduces into various literatures. First, demonstrating the value of interdisciplinary research, it brings the psychological construct of self-regulation failure in relation to two philosophical conceptions of weakness of will and helps to see that the notion of self-regulation failure is not synonymous with either but instead captures a wider variety of agentic failings and lays bare more of the complexity of the causal mechanisms underlying these failings. Second, the dissertation provides a more precise understanding of complacency and its relation to automation bias, hoping to move the state of the literature on automation bias and

automation-related complacency beyond the conceptual confusions that currently dominate it. Third, meeting the first aim mentioned in the Introduction, the dissertation offers a new, comprehensive definition of e-coaching systems that helps to understand how e-coaching systems by their level of sophistication and independence differ from other (persuasive) self-regulation facilitators. The value of this latter contribution is already evidenced by the uptake of the definition in recent literature (e.g., Beinema, Op Den Akker, and Hermens, 2018; Ochoa and Gutierrez, 2018).

The second contribution the dissertation makes, meeting its second aim, is the disentanglement of various ethical concerns that arise with the widespread adoption of e-coaching technologies. By bringing into view the ethical landscape surrounding e-coaching systems, the hope is to help structure future discussions in a variety of contexts about the use and regulation of e-coaching systems. The inventory of concerns it provides is not exhaustive or definitive, but is meant rather as a point of departure from which to examine, discuss and further develop the various distinct concerns, their relationships to one another, and potential mitigation strategies.

The third contribution the dissertation makes, meeting its third aim, is the philosophical analysis it offers of the ways in which self-regulation support systems run the risk of negatively affecting people's personal autonomy by facilitating complacency in aspects of people's practical reasoning that undermines both their synchronic and diachronic self-governance. Separating the empirical claim (Heightened Risk of Complacency) from the conceptual claim (Implication for Self-Governance) contributes to a clearer understanding of the complexity of the concern, and highlights the need for an interdisciplinary research approach to dealing with this concern.

The fourth and final contribution this dissertation makes, meeting its fourth aim, is the set of "anti-complacency" recommendations it offers about ensuring ongoing consent, revealing the reasoning behind suggestions, increasing user awareness about system fallibility, offering reassessment opportunities, and promoting reflection on suggestions. The recommendations are accompanied with a set of diverse research questions, further underlining the relevance of interdisciplinary work in relation to e-coaching systems. Despite the open questions, though, the recommendations should be concrete enough to serve

as initial guidelines for implementing a set of first-generation anti-complacency measures, even if these will require careful evaluation and further improvement over time.

Overall, I hope the view I have set forth in this dissertation on how the strengths of e-coaching systems are potentially related to pitfalls for users of such systems, is neither discouraging nor disheartening for those who see a bright future for e-coaching technologies. For it exactly the point of works like this one to evoke discussion about how to actively shape a bright future, rather than passively await social, political and cultural disruption that is driven by what is technologically possible instead of what is desirable. Thus, this dissertation as a whole, including the recommendations, should serve as input to the aforementioned ongoing discussions about the further development of e-coaching technologies, paving the way for further, future contributions to the responsible innovation of effective e-coaching systems that fend off complacency and are respectful of people's personal autonomy.

References

Aarts, E. and B. De Ruyter (2009). New research perspectives on Ambient Intelligence. *Journal of Ambient Intelligence and Smart Environments 1*(1), 5–14.

Abdul, A., J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 582. ACM.

Adams, F. and K. Aizawa (2001). The Bounds of Cognition. *Philosophical psychology 14*(1), 43–64.

Adriaanse, M. A., G. Oettingen, P. M. Gollwitzer, E. P. Hennes, D. T. D. De Ridder, and J. B. F. De Wit (2010). When planning is not enough: Fighting unhealthy snacking habits by mental contrasting with implementation intentions (MCII). *European Journal of Social Psychology 40*, 1277–1293.

Ainslie, G. (2001). *Breakdown of Will*. Cambridge, UK: Cambridge University Press.

Aizawa, K. (2018). Extended Cognition, Trust and Glue, and Knowledge. In J. A. Carter, A. Clark, J. Kallestrup, S. Orestis Palermos, and D. Pritchard (Eds.), *Extended Epistemology*, Chapter 3, pp. 64–78. Oxford University Press.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes 50*(2), 179–211.

Alberdi, E., L. Strigini, A. A. Povyakalo, and P. Ayton (2009). Why Are People's Decisions Sometimes Worse with Computer Support? In B. Buth, G. Rabe, and T. Seyfarth (Eds.), *Computer Safety, Reliability, and Security*, Volume 5775 of *Lecture Notes in Computer Science*, pp. 18–31. Springer Berlin Heidelberg.

Aminikhanghahi, S., R. Fallahzadeh, M. Sawyer, D. J. Cook, and L. B. Holder (2017). Thyme: Improving Smartphone Prompt Timing Through Activity Awareness. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 315–322. IEEE.

Anderson, H. and S. Swim (1995). Supervision as Collaborative Conversation: Connecting the Voices of Supervisor and Supervisee. *Journal of Systemic Therapies 14*(2), 1–13.

Anderson, J. and A. Honneth (2005). Autonomy, Vulnerability, Recognition,

and Justice. In *Autonomy and the challenges to liberalism: New essays*, pp. 127–149. Cambridge, UK: Cambridge University Press.

Anderson, J. H. (2009). Autonomy Gaps as a Social Pathology: Ideologiekritik Beyond Paternalism. In R. Forst, M. Hartmann, R. Jaeggi, and M. Saar (Eds.), *Sozialphilosophie und Kritik*. Suhrkamp.

Anderson, J. H. (2013). Autonomy. In H. LaFollette, J. Deigh, and S. Stroud (Eds.), *International Encyclopedia of Ethics*. Hoboken, NJ: Wiley-Blackwell.

Anderson, J. H. and B. A. Kamphorst (2014). Ethics of e-coaching: Implications of employing pervasive computing to promote healthy and sustainable lifestyles. In *Proceedings of the third IEEE International Workshop on the Social Implications of Pervasive Computing for Sustainable Living (SIPC 2014), in conjunction with the Twelfth IEEE International Conference on Pervasive Computing and Communications (PerCom 2014)*, pp. 351–356. IEEE Computer Society Press.

Anderson, J. H. and B. A. Kamphorst (2015). Should Uplifting Music and Smart Phone Apps Count as Willpower Doping? The Extended Will and the Ethics of Enhanced Motivation. *American Journal of Bioethics: Neuroscience 6*(1), 35–37.

Anscombe, G. E. M. (1963). *Intention* (Second ed.). Oxford: Basil Blackwell. (First edition 1957).

Ariely, D. (2008). *Predictably Irrational*. New York, NY: HarperCollins.

Austin, J. T. and J. B. Vancouver (1996). Goal constructs in psychology: Structure, process, and content. *Psychological bulletin 120*(3), 338.

Aydin, C. (2013). The artifactual mind: Overcoming the 'inside-outside' dualism in the extended mind thesis and recognizing the technological dimension of cognition. *Phenomenology and the Cognitive Sciences*, 1–22.

Bagheri, N. and G. A. Jamieson (2004). Considering Subjective Trust and Monitoring Behavior in Assessing Automation-Induced "Complacency". In D. A. Vincenzi, M. Mouloua, and P. A. Hancock (Eds.), *Human Performance, Situation Awareness and Automation: Current Research and Trends*, Volume II, pp. 54–59.

Bahner, J. E., A.-D. Hüper, and D. Manzey (2008). Misuse of Automated Decision Aids: Complacency, Automation Bias and the Impact of Training Experience. *International Journal of Human-Computer Studies 66*(9), 688–699.

Bamberg, S. (2002). Effects of Implementation Intententions on the Actual Performance of New Environmentally Friendly Behaviours — Results of Two Field Experiments. *Journal of Environmental Psychology 22*(4), 399–411.

Bandura, A. (1991). Social Cognitive Theory of Self-Regulation. *Organizational behavior and human decision processes 50*(2), 248–287.

Bandura, A. (1997). *Self-efficacy: The Exercise of Control*. New York, NY: Freeman.

Bandura, A. (2005). The Primacy of Self-Regulation in Health Promotion. *Applied Psychology 54*(2), 245–254.

Bargh, J. A., P. M. Gollwitzer, A. Lee-Chai, K. Barndollar, and R. Trötschel (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of personality and social psychology 81*(6), 1014.

Bargh, J. A. and E. L. Williams (2006). The Automaticity of Social Life. *Current directions in psychological science 15*(1), 1–4.

Barlyn, S. (2018). John hancock will only sell interactive life insurance with fitness data tracking. `https://www.insurancejournal.com/news/national/2018/09/19/501747.htm`. Accessed: 2019-09-11.

Barocas, S. and H. Nissenbaum (2014). Big Data's End Run Around Anonymity and Consent. In J. Lane, V. Stodden, S. Bender, and H. Nissenbaum (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, pp. 44–75. Cambridge University Press, NY.

Barutchu, A., O. Carter, R. Hester, and N. Levy (2013). Strength in Cognitive Self-Regulation. *Frontiers in Psychology 4*, 174.

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of personality and social psychology 46*(3), 610.

Baumeister, R. F., M. Gailliot, C. N. DeWall, and M. Oaten (2006). Self-regulation and personality: How interventions increase regulatory success, and how depletion moderates the effects of traits on behavior. *Journal of Personality 74*(6), 1773–1802.

Baumeister, R. F. and T. F. Heatherton (1996). Self-Regulation Failure: An Overview. *Psychological inquiry 7*(1), 1–15.

Baumeister, R. F., T. F. Heatherton, and D. M. Tice (1994). *Losing Control: How and Why People Fail at Self-Regulation*. San Diego, CA: Academic Press.

Baumeister, R. F. and K. D. Vohs (2007). Self-Regulation, Ego Depletion, and Motivation. *Social and Personality Psychology Compass 1*(1), 115–128.

Beauchamp, T. L. (2005). Who Deserves Autonomy, and Whose Autonomy Deserves Respect. In J. S. Taylor (Ed.), *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*, pp. 310–329. Cambridge University Press.

Beinema, T., H. Op Den Akker, and H. Hermens (2018). Creating an Articial Coaching Engine for Multi-domain Conversational Coaches in eHealth Applications. In *Proceedings of ACM Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications (ICA-HoGeCa2018)*.

Berdichevsky, D. and E. Neuenschwander (1999). Toward an Ethics of Persuasive Technology. *Communications of the ACM 42*(5), 51–58.

Beun, R. J., R. Ahn, F. Griffioen Both, S. Fitrianie, and J. Lancee (2014). Modeling Interaction in Automated E-Coaching — A Case from Insomnia Therapy. In H. Lounis and D. Josyula (Eds.), *Proceedings of the Sixth International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2014)*. IARIA.

Beun, R. J., W. P. Brinkman, S. Fitrianie, F. Griffioen-Both, C. Horsch, J. Lancee, and A. G. L. Spruit (2016). Adherence in Automated e-Coaching: a Case from Insomnia Therapy. In *Persuasive Technology. Proceedings of the 11th Int. Conf. PERSUASIVE 2016*, Volume 9638 of *LNCS*, Salsburg, Austria.

Bidargaddi, N., D. Almirall, S. Murphy, I. Nahum-Shani, M. Kovalcik, T. Pituch, H. Maaieh, and V. Strecher (2018). To Prompt or Not to Prompt? A Microrandomized Trial of Time-Varying Push Notifications to Increase Proximal Engagement With a Mobile Health App. *JMIR mHealth and uHealth 6*(11), e10123.

Biran, O. and C. Cotton (2017). Explanation and Justification in Machine Learning: A Survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Volume 8, pp. 1.

Boekaerts, M., P. R. Pintrich, and M. Zeidner (Eds.) (2000). *Handbook of Self-Regulation*. Academic Press.

Boh, B., L. H. J. M. Lemmens, A. Jansen, C. Nederkoorn, V. Kerkhofs, G. Spanakis, G. Weiss, and A. Roefs (2016). An Ecological Momentary Intervention for weight loss and healthy eating via smartphone and Internet: Study protocol for a randomised controlled trial. *Trials 17*(1), 1.

Bohn, J., V. Coroamă, M. Langheinrich, F. Mattern, and M. Rohs (2005). Social, Economic, and Ethical Implications of Ambient Intelligence and Ubiquitous Computing. In W. Weber, J. Rabaey, and E. Aarts (Eds.), *Ambient Intelligence*, pp. 5–29. Springer Berlin Heidelberg.

Bratman, M. E. (1987). *Intentions, Plans and Practical Reason*. Harvard University Press.

Bratman, M. E. (2007). *Structures of Agency: Essays*. New York, NY: Oxford University Press.

Bratman, M. E. (2014). *Shared Agency. A planning Theory of Acting Together*. New York, NY: Oxford University Press.

Bratman, M. E. (2018). *Planning, Time, and Self-Governance: Essays in Practical Rationality*. Oxford University Press.

Bratman, M. E., D. J. Israel, and M. E. Pollack (1988). Plans and Resource-Bounded Practical Reasoning. *Computational intelligence 4*(3), 349–355.

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in cognitive sciences 16*(2), 106–113.

Braver, T. S., M. K. Krug, K. S. Chiew, W. Kool, J. A. Westbrook, N. J. Clement, R. A. Adcock, D. M. Barch, M. M. Botvinick, C. S. Carver, R. Cools, R. Custers, A. Dickinson, C. S. Dweck, A. Fishbach, P. M. Gollwitzer, T. M. Hess, D. M. Isaacowitz, M. Mather, K. Murayama, L. Pessoa, G. R. Samanez-Larkin, and L. H. Somerville (2014). Mechanisms of Motivation-Cognition Interaction: Challenges and Opportunities. *Cognitive, Affective, & Behavioral Neuroscience 14*(2), 443–472.

Brodscholl, J. C., H. Kober, and E. T. Higgins (2007). Strategies of self-

regulation in goal attainment versus goal maintenance. *European Journal of Social Psychology 37*(4), 628–648.

Broersen, J., M. Dastani, and L. Van der Torre (2005). Beliefs, obligations, intentions, and desires as components in an agent architecture. *International Journal of Intelligent Systems 20*(9), 893–919.

Brown, I. and A. A. Adams (2007). The ethical challenges of ubiquitous healthcare. *International Review of Information Ethics 8*, 53–60.

Brustoloni, J. C. (1991). Autonomous Agents: Characterization and Requirements. Carnegie Mellon Technical Report CMU-CS-91-204, Carnegie Mellon University.

Burgess, A., H. Cappelen, and D. Plunkett (Eds.) (2019). *Conceptual Engineering and Conceptual Ethics*. Oxford, UK: Oxford University Press.

Buxton, O. M. and E. Marcelli (2010). Short and long sleep are positively associated with obesity, diabetes, hypertension, and cardiovascular disease among adults in the United States. *Social Science & Medicine 71*(5), 1027–1036.

Carter, J. A., A. Clark, and S. O. Palermos (2018). New Humans? Ethics, Trust, and the Extended Mind. pp. 331–351.

Carter, J. A. and S. O. Palermos (2016). Is Having Your Computer Compromised a Personal Assault? The Ethics of Extended Cognition. *Journal of the American Philosophical Association*.

Carver, C. S., S. L. Johnson, and J. Joormann (2009). Two-Mode Models of Self-Regulation as a Tool for Conceptualizing Effects of the Serotonin System in Normal Behavior and Diverse Disorders. *Current Directions in Psychological Science 18*(4), 195–199.

Carver, C. S. and M. F. Scheier (1981). *Attention and Self-Regulation: A Control-Theory Approach to Human Behavior*. New York, NY: Springer-Verlag.

Carver, C. S. and M. F. Scheier (1982). Control Theory: A Useful Conceptual Framework for Personality–Social, Clinical, and Health Psychology. *Psychological bulletin 92*(1), 111.

Cave, E. M. (2007). What's Wrong with Motive Manipulation? *Ethical Theory and Moral Practice 10*(2), 129–144.

Chaiken, S. and Y. Trope (1999). *Dual-Process Theories in Social Psychology*. Guilford Press.

Christensen, A. J., P. J. Moran, J. S. Wiebe, S. L. Ehlers, and W. J. Lawton (2002). Effect of a Behavioral Self-Regulation Intervention on Patient Adherence in Hemodialysis. *Health Psychology 21*(4), 393–397.

Christman, J. (2004). Relational Autonomy, Liberal Individualism, and the Social Constitution of Selves. *Philosophical Studies 117*(1-2), 143–164.

Christman, J. (2009). *The Politics of Persons: Individual Autonomy and Socio-Historical Selves*. Cambridge, UK: Cambridge University Press.

Christman, J. and J. H. Anderson (Eds.) (2005). *Autonomy and the Challenges to Liberalism: New Essays*. New York, NY: Cambridge University Press.

Clare, A. S., M. L. Cummings, and N. P. Repenning (2015). Influencing Trust for Human–Automation Collaborative Scheduling of Multiple Unmanned Vehicles. *Human Factors: The Journal of the Human Factors and Ergonomics Society 57*(7), 1208–1218.

Clark, A. (2003). *Natural-Born Cyborgs: Mind, Technologies, and the Future of Human Intelligence*. Oxford, UK: Oxford University Press.

Clark, A. (2011). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York, NY: Oxford University Press.

Clark, A. and D. Chalmers (1998). The Extended Mind. *Analysis 58*(1), 7–19.

Coiera, E. (2015). Technology, cognition and error. *BMJ Quality & Safety 24*(7), 417–422.

Conly, S. (2013). *Against Autonomy: Justifying Coercive Paternalism*. Cambridge, UK: Cambridge University Press.

Cooke, R. and P. Sheeran (2004). Moderation of cognition-intention and cognition-behaviour relations: A meta-analysis of properties of variables from the theory of planned behaviour. *British Journal of Social Psychology 43*(2), 159–186.

Crisp, R. and C. Cowton (1994). Hypocrisy and Moral Seriousness. *American Philosophical Quarterly 31*(4), 343–349.

Cummings, M. L. (2004). Automation Bias in Intelligent Time Critical Decision Support Systems. In *Proceedings of AIAA 3rd Intelligent Systems Conference*, pp. 2004–6313. AIAA.

Custers, B. (2016). Click here to consent forever: Expiry dates for informed consent. *Big Data & Society 3*(1), 2053951715624935.

Davidson, D. (1963). Actions, Reasons, and Causes. *Journal of Philosophy* (60), 685–700. Reprinted in 'The Essential Davidson', Oxford University Press (2006). Citations refer to this edition.

Davidson, D. (1973). Freedom to Act. In T. Honderich (Ed.), *Essays on Freedom of Action*. Routledge.

De Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen, and V. D. Blondel (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports 3*, 1376.

De Ridder, D. T. D., E. De Vet, F. M. Stok, M. A. Adriaanse, and J. B. F. De Wit (2013). Obesity, overconsumption and self-regulation failure: The unsung role of eating appropriateness standards. *Health Psychology Review 7*(2), 146–165.

De Ridder, D. T. D. and M. Gillebaart (2017). Lessons learned from trait self-control in well-being: Making the case for routines and initiation as important components of trait self-control. *Health psychology review 11*(1), 89–99.

De Ridder, D. T. D., G. Lensvelt-Mulders, C. Finkenauer, F. M. Stok, and R. F. Baumeister (2012). Taking Stock of Self-Control: A Meta-Analysis of How

Trait Self-Control Relates to a Wide Range of Behaviors. *Personality and Social Psychology Review 16*(1), 76–99.

Dennett, D. C. (1987). True Believers. In *The Intentional Stance*, Chapter 2. Cambridge, MA: MIT Press.

Dennett, D. C. (2014). Commentary on Kamphorst and Kalis. *Tijdschrift voor Filosofie 76*, 583.

Donovan, D. A. and S. M. Wildman (1980). Is the Reasonable Man Obsolete: A Critical Perspective on Self-Defense and Provocation. *Loyola of Los Angeles Law Review 14*, 435.

Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. *Management review 70*(11), 35–36.

Dunbar-Jacob, J. and E. Schlenk (2000). Patient Adherence to Treatment Regimens. In A. Baum, T. Revenson, and J. Singer (Eds.), *Handbook of health psychology*, pp. 571–580. Mahwah, NJ: Erlbaum.

Dworkin, G. (1976). Autonomy and Behavior Control. *Hastings Center Report 6*(1), 23–28.

Dworkin, G. (2015). The Nature of Autonomy. *Nordic Journal of Studies in Educational Policy 2015*(2), 28479.

European Court of Justice (Grand Chamber) (2019). Judgment of 24 September 2019. Google v. CNIL, C507/17, ECLI:EU:C:2019:772.

Faden, R. R. and T. L. Beauchamp (1986). *A History and Theory of Informed Consent*. Oxford, UK: Oxford University Press.

Fan, M. and Y. Jin (2014). Obesity and Self-control: Food Consumption, Physical Activity, and Weight-loss Intention. *Applied Economic Perspectives and Policy 36*(1), 125–145.

Feinberg, J. (1986). *The Moral Limits of the Criminal Law. Volume 3, Harm to Self*. Oxford University Press.

Finkelstein, E. A., I. C. Fiebelkorn, and G. Wang (2003). National Medical Spending Attributable To Overweight And Obesity: How Much, And Who's Paying? *Health Affairs*.

Fischer, J. M. (2004). Responsibility and Manipulation. *The Journal of Ethics 8*(2), 145–177.

Fishbach, A. and W. Hofmann (2015). Nudging self-control: A smartphone intervention of temptation anticipation and goal resolution improves everyday goal progress. *Motivation Science 1*(3), 137.

Fiske, S. T. and S. E. Taylor (1991). *Social Cognition* (2nd ed.). New York, NY: McGraw-Hill.

Flanagin, A. J. and M. M. J. (2008). Digital Media and Youth: Unparalleled Opportunity and Unprecedented Responsibility. In M. J. Metzger and A. J. Flanagin (Eds.), *Digital Media, Youth, and Credibility*, Series on Digital Media and Learning, pp. 5–28. Cambridge, MA: The MIT Press.

Fogg, B. J. (2003). *Persuasive Technology: Using Computers to Change What We*

Think and Do. San Fransisco: Morgan Kaufmann Publishers.

Frankfurt, H. (1982). The Importance of What We Care About. *Synthese 53*, 257–272.

Frankfurt, H. (1987). Identification and Wholeheartedness. In F. D. Schoeman (Ed.), *Responsiblity, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge University Press.

Frankfurt, H. (1992). The Faintest Passion. *Proceedings and Addresses of the American Philosophical Association 66*(3), 5–16.

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy 68*(1), pp. 5–20.

Frankfurt, H. G. (1978). The Problem of Action. *American philosophical quarterly 15*(2), 157–162.

Frankish, K. (2010). Dual-Process and Dual-System Theories of Reasoning. *Philosophy Compass 5*(10), 914–926.

Franklin, S. and A. Graesser (1997). Is It an agent, or just a program?: A taxonomy for autonomous agents. In J. P. Müller, M. J. Wooldridge, and N. R. Jennings (Eds.), *Intelligent Agents III Agent Theories, Architectures, and Languages*, Volume 1193 of *Lecture Notes in Computer Science*.

Frey, J. (2013). AdAPT–A Dynamic Approach for Activity Prediction and Tracking for Ambient Intelligence. In *Intelligent Environments (IE), 2013 9th International Conference on*, pp. 254–257. IEEE.

Friese, M. and W. Hofmann (2009). Control me or I will control you: Impulses, trait self-control, and the guidance of behavior. *Journal of Research in Personality 43*(5), 795–805.

Fujita, K. (2008). Seeing the Forest Beyond the Trees: A Construal-Level Approach to Self-Control. *Social and Personality Psychology Compass 2*(3), 1475–1496.

Fujita, K. (2011). On Conceptualizing Self-Control as More Than the Effortful Inhibition of Impulses. *Personality and Social Psychology Review 15*(4), 352–366.

Gallagher, S. (2013). The socially extended mind. *Cognitive Systems Research 25*, 4–12.

Gallagher, S. and M. Bower (2014). Making enactivism even more embodied. *Avant: Trends in Interdisciplinary Studies* (2), 232–247.

Garg, A. X., N. J. Adhikari, H. McDonald, M. Rosas-Arellano, P. J. Devereaux, J. Beyene, J. Sam, and B. Haynes (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *Journal of the American Medical Association 293*(10), 1223–1238.

Giraffa, L. M. M. and R. M. Viccari (1998). The Use of Agents Techniques on Intelligent Tutoring Systems. In *Computer Science, 1998. SCCC'98. XVIII International Conference of the Chilean Society of*, pp. 76–83. IEEE.

Goddard, K., A. Roudsari, and J. C. Wyatt (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association 19*(1), 121–127.

Goddard, K., A. Roudsari, and J. C. Wyatt (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics 83*(5), 368–375.

Goldman, A. I. (1970). *Theory of Human Action*. Princeton University Press.

Gollwitzer, P. M. (1993, Oct). Goal Achievement: The Role of Intentions. *European Review of Social Psychology 4*, 141–185.

Gollwitzer, P. M. (1996). The Volitional Benefits of Planning. In *The Psychology of Action: Linking cognition and motivation to behavior*, pp. 287–312. New York: Guilford.

Gollwitzer, P. M. and P. Sheeran (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology 38*, 69–119.

Grant, A. M. and D. R. Stober (2006). Introduction. In D. R. Stober and A. M. Grant (Eds.), *Evidence Based Coaching Handbook: Putting Best Practices to Work for Your Clients*. New York, NY: Wiley.

Greenblum, J. and R. Hubbard (2019). The common rule's "reasonable person" standard for informed consent. *Bioethics 33*(2), 274–277.

Gulz, A. (2004). Benefits of Virtual Characters in Computer Based Learning Environments: Claims and Evidence. *International Journal of Artificial Intelligence in Education (IJAIED) 14*, 313–334.

Haghbin, M., A. McCaffrey, and T. A. Pychyl (2012). The Complexity of the Relation Between Fear of Failure and Procrastination. *Journal of Rational-Emotive & Cognitive-Behavior Therapy 30*(4), 249–263.

Hall, P. and N. Gill (2018). *An Introduction to Machine Learning Interpretability*. O'Reilly Media, Incorporated.

Harbers, M., J. M. Bradshaw, M. Johnson, P. Feltovich, K. Van den Bosch, and J.-J. Meyer (2012). Explanation in Human-Agent Teamwork. In S. Cranefield, M. B. Van Riemsdijk, J. Vázquez-Salceda, and P. Noriega (Eds.), *Coordination, Organizations, Institutions, and Norms in Agent System VII: COIN 2011 International Workshops, COIN@AAMAS 2011, Taipei, Taiwan, May 3, 2011, COIN@WI-IAT 2011, Lyon, France, August 22, 2011, Revised Selected Papers*, pp. 21–37. Berlin, Heidelberg: Springer Berlin Heidelberg.

Harbers, M., K. Van den Bosch, and J.-J. Meyer (2010). A Methodology for Developing Self-explaining Agents for Virtual Training. In M. Dastani, A. El Fallah Segrouchni, J. Leite, and P. Torroni (Eds.), *Languages, Methodologies, and Development Tools for Multi-Agent Systems: Second International Workshop, LADS 2009, Torino, Italy, September 7-9, 2009, Revised Selected Papers*, pp. 168–182. Berlin, Heidelberg: Springer Berlin Heidelberg.

Harbo, T. (2010). The Function of the Proportionality Principle in EU Law.

European Law Journal 16(2), 158–185.

Harkin, B., T. L. Webb, B. P. I. Chang, A. Prestwich, M. Conner, Y. Kellar, I. Benn, and P. Sheeran (2015). Does Monitoring Goal Progress Promote Goal Attainment? A Meta-Analysis of the Experimental Evidence. *Psychological Bulletin*. Online first.

Harman, G. (1976). Practical Reasoning. *The Review of Metaphysics 29*(3), 431–463.

Harman, G. (1986). *Change in View: Principles of Reasoning*. Cambridge, MA: The MIT Press.

Harman, G. (1999). Practical Reasoning. In *Reasoning, Meaning and Mind*, Chapter 2. Oxford University Press.

Harrison, Y. and J. A. Horne (2000). Sleep loss and temporal memory. *The Quarterly Journal of Experimental Psychology: Section A 53*(1), 271–279.

Hayes, E. and K. A. Kalmakis (2007). From the sidelines: Coaching as a nurse practitioner strategy for improving health outcomes. *Journal of the American Academy of Nurse Practitioners 19*(11), 555–562.

Heath, J. and J. H. Anderson (2010). Procrastination and the Extended Will. In *The Thief of Time: Philosophical Essays on Procrastination*, pp. 233–252. Oxford University Press.

Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences 14*(3), 577–598.

Heersmink, R. (2017). Extended mind and cognitive enhancement: Moral aspects of cognitive artifacts. *Phenomenology and the Cognitive Sciences 16*(1), 17–32.

Hernandez, J. T. and R. J. Diclemente (1992). Self-control and ego identity development as predictors of unprotected sex in late adolescent males. *Journal of Adolescence 15*(4), 437–447.

Hofmann, W., B. J. Schmeichel, and A. D. Baddeley (2012). Executive functions and self-regulation. *Trends in cognitive sciences 16*(3), 174–180.

Holland, R. W., H. Aarts, and D. Langendam (2006). Breaking and creating habits on the working floor: A field-experiment on the power of implementation intentions. *Journal of Experimental Social Psychology 42*(6), 776–783.

Holton, R. (1999). Intention and Weakness of Will. *The Journal of Philosophy 96*(5), 241–262.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford University Press.

Hooft, E. A. J. v., M. P. Born, T. W. Taris, H. van der Flier, and R. W. B. Blonk (2005). Bridging the gap between intentions and behavior: Implementation intentions, action control, and procrastination. *Journal of Vocational Behavior* (66), 238–256.

Horsch, C. H. G., J. Lancee, F. Griffioen-Both, A. G. L. S. Spruit, S. Fitrianie, M. A. Neerincx, R. J. Beun, and W.-P. Brinkman (2017). Mobile Phone-Delivered Cognitive Behavioral Therapy for Insomnia: A Randomized

Waitlist Controlled Trial. *Journal of medical Internet research 19*(4), e70.

Hoyle, R. H. (2006). Personality and Self-Regulation: Trait and Information-Processing Perspectives. *Journal of Personality 74*(6), 1507–1526.

Hurn, J., I. Kneebone, and M. Cropley (2006). Goal setting as an outcome measure: A systematic review. *Clinical rehabilitation 20*(9), 756–772.

Hutto, D. and E. Myin (2013). *Radicalizing Enactivism: Basic Minds Without Content*. Cambridge, MA: MIT Press.

Inzlicht, M. and B. J. Schmeichel (2012). What is Ego Depletion? Toward a Mechanistic Revision of the Resource Model of Self-Control. *Perspectives on Psychological Science 7*(5), 450–463.

Ives, Y. (2008). What is 'Coaching'? An Exploration of Conflicting Paradigms. *International Journal of Evidence Based Coaching and Mentoring 6*.

Joffe, S., E. F. Cook, P. D. Cleary, J. W. Clark, and J. C. Weeks (2001). Quality of informed consent in cancer clinical trials: A cross-sectional survey. *The Lancet 358*(9295), 1772–1777.

John, P., G. Smith, and G. Stoker (2009). Nudge nudge, think think: Two strategies for changing civic behaviour. *The Political Quarterly 80*(3), 361–370.

Kalis, A. (2011). *Failures of Agency: Irrational Behavior and Self-Understanding*. Rowman & Littlefield Publishers.

Kamphorst, B. A. (2017). E-Coaching Systems: What They Are, and What They Aren't. *Personal and Ubiquitous Computing 21*, 625–632.

Kamphorst, B. A., J. H. Anderson, and M. V. Dignum (in prep.). An Unobtrusive Logging and Intervention Tool for Procrastination Studies: Introducing the "ii-app".

Kamphorst, B. A. and A. Kalis (2015). Why option generation matters for the design of autonomous e-coaching systems. *AI & Society 30*, 77–88.

Kamphorst, B. A., M. C. A. Klein, and A. Van Wissen (2014a). Autonomous e-Coaching in the Wild: Empirical Validation of a Model-based Reasoning System. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, Richland, SC, pp. 725–732. International Foundation for Autonomous Agents and Multiagent Systems.

Kamphorst, B. A., M. C. A. Klein, and A. Van Wissen (2014b). Human Involvement in E-Coaching: Effects on Effectiveness, Perceived Influence and Trust. In H. S. Park, A. A. Salah, Y. J. Lee, L.-P. Morency, Y. Sheikh, and R. Cucchiara (Eds.), *Human Behavior Understanding*, Volume 8749 of *Lecture Notes in Computer Science*, pp. 16–29. Springer International Publishing.

Kamphorst, B. A., S. Nauts, and E.-M. Blouin-Hudon (2017). Introducing a Continuous Measure of Future Self-Continuity. *Social Science Computer Review 35*(3).

Kamphorst, B. A., S. Nauts, D. T. D. De Ridder, and J. H. Anderson (2018). Too Depleted to Turn In: The Relevance of End-of-the-Day Resource Depletion

for Reducing Bedtime Procrastination. *Frontiers in Psychology 9*, 252.

Kane, R. (1998). *The Significance of Free Will*. Oxford, UK: Oxford University Press.

Kanfer, R. (1990). Motivation and individual differences in learning: An integration of developmental, differential and cognitive perspectives. *Learning and Individual Differences 2*(2), 221–239.

Karoly, P. (1993). Mechanisms of self-regulation: A systems view. *Annual review of psychology 44*(1), 23–52.

Karppinen, P. and H. Oinas-Kukkonen (2013). Three Approaches to Ethical Considerations in the Design of Behavior Change Support Systems. In S. Berkovsky and J. Freyne (Eds.), *Persuasive Technology*, Volume 7822 of *Lecture Notes in Computer Science*, pp. 87–98. Springer Berlin Heidelberg.

Kawall, J. (2006). On Complacency. *American Philosophical Quarterly 43*(4), 343–355.

Kelsen, H. (1967). *The Pure Theory of Law*. Los Angeles, CA: University of California Press. Translated from the Second German edition by Max Knight.

Kidwell, B. D., W. D. Miller, and R. Parasuraman (2014). Automation Complacency: Using Non-Invasive Brain Stimulation to Change Attention Allocation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 58*(1), 240–244.

Kjaersgaard, T. (2015). Enhancing Motivation by Use of Prescription Stimulants: The Ethics of Moral Enhancement. *American Journal of Bioethics: Neuroscience 6*(1), 4–10.

Klein, M., N. Mogles, and A. Van Wissen (2013). An Intelligent Coaching System for Therapy Adherence. *IEEE pervasive computing 12*(3), 22–30.

Klein, M. C. A., N. M. Mogles, and A. Van Wissen (2014). Intelligent mobile support for therapy adherence and behavior change. *Journal of Biomedical Informatics 51*, 137–151.

Knobe, J. (2007). Experimental Philosophy. *Philosophy Compass 2*(1), 81–92.

Kool, L., J. Timmer, and R. Van Est (Eds.) (2014). *Eerlijk advies: De opkomst van de e-coach*. Den Haag, Netherlands: Rathenau Instituut.

Kopp, C. B. (1982). Antecedents of Self-Regulation: A Developmental Perspective. *Developmental psychology 18*(2), 199.

Kroese, F. M., D. T. D. De Ridder, C. Evers, and M. A. Adriaanse (2014). Bedtime procrastination: Introducing a new area of procrastination. *Frontiers in Psychology 5*(611).

Kupfer, J. (1987). Privacy, Autonomy, and Self-Concept. *American Philosophical Quarterly 24*(1), 81–89.

Lamond, G. (2000). The Coerciveness of Law. *Oxford Journal of Legal Studies 20*(1), 39–62.

Latham, G. P. and E. A. Locke (1991). Self-Regulation through Goal Setting. *Organizational behavior and human decision processes 50*(2), 212–247.

Levy, N. (2011). Resisting 'Weakness of the Will'. *Philosophy and Phenomenological Research 82*(1), 134–155.

Levy, N. (2012). Ecological Engineering: Reshaping Our Environments to Achieve Our Goals. *Philosophy & Technology 25*, 589–604.

Little, B. R. (2014). *Me, Myself, and Us: The Science of Personality and the Art of Well-Being*. New York, NY: PublicAffairs.

Locke, E. A. and G. P. Latham (2002). Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-Year Odyssey. *American psychologist 57*(9), 705.

Locke, E. A., K. N. Shaw, L. M. Saari, and G. P. Latham (1981). Goal Setting and Task Performance: 1969–1980. *Psychological bulletin 90*(1), 125.

Logan, G. D. and W. B. Cowan (1984). On the Ability to Inhibit Thought and Action: A Theory of an Act of Control. *Psychological review 91*(3), 295.

Lowe, E. J. (2008). *Personal Agency: The Metaphysics of Mind and Action*. New York, NY: Oxford University Press.

Lucas, J. R. (1966). *The Principles of Politics*. Oxford: Clarendon Press.

MacIntyre, A. (1981). *After Virtue*. Notre Dame, IN: University of Notre Dame Press.

Mackenzie, C. (2008). Relational Autonomy, Normative Authority and Perfectionism. *Journal of Social Philosophy 39*(4), 512–533.

MacLeod, L. (2012). Making SMART goals smarter. *Physician executive 38*(2), 68–72.

Mann, T., D. T. D. De Ridder, and K. Fujita (2013). Self-Regulation of Health Behavior: Social Psychological Approaches to Goal Setting and Goal Striving. *Health Psychology 32*(5), 487.

Martani, A., D. Shaw, and B. S. Elger (2019). Stay fit or get bit–ethical issues in sharing health data with insurers' apps. *Swiss medical weekly 149*(2526).

Masicampo, E. J. and R. F. Baumeister (2011). Consider It Done! Plan Making Can Eliminate the Cognitive Effects of Unfulfilled Goals. *Journal of personality and social psychology 101*(4), 667.

May, J. and R. Holton (2012). What in the world is weakness of will? *Philosophical Studies 157*(3), 341–360.

Mayer, J. D. and Y. N. Gaschke (1988). The experience and meta-experience of mood. *Journal of personality and social psychology 55*(1), 102.

Mayer, R. E. and C. S. DaPra (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied 18*, 239–252.

McGuirl, J. M. and N. B. Sarter (2003). How are we doing?: Presenting System Confidence Information to Support Trust Calibration and Adaptive Function Allocation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 47*(3), 538–542.

McIntyre, A. (2006). What Is Wrong with Weakness of Will? *The Journal of*

Philosophy 103(6), pp. 284–311.

Meacham, S., G. Isaac, D. Nauck, and B. Virginas (2019). Towards Explainable AI: Design and Development for Explanation of Machine Learning Predictions for a Patient Readmittance Medical Application. In *Intelligent Computing-Proceedings of the Computing Conference*, pp. 939–955. Springer.

Mele, A. (1995). *Autonomous Agents: From Self-Control to Autonomy*. New York, NY: Oxford University Press.

Mele, A. (2003). *Motivation and Agency*. New York, NY: Oxford University Press.

Mele, A. (2010). Weakness of Will and Akrasia. *Philosophical Studies 150*(3), 391–404.

Mele, A. R. (2012). *Backsliding*. New York, NY: Oxford University Press.

Menary, R. (2006). Attacking the Bounds of Cognition. *Philosophical Psychology 19*(3), 329–344.

Menary, R. (2010a). Dimensions of mind. *Phenomenology and the Cognitive Sciences 9*(4), 561–578.

Menary, R. (2010b). Introduction to the special issue on 4E cognition. *Phenomenology and the Cognitive Sciences 9*(4), 459–463.

Metcalfe, J. and W. Mischel (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological review 106*(1), 3.

Michie, S., S. Ashford, F. F. Sniehotta, S. U. Dombrowski, A. Bishop, and D. P. French (2011). A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: The CALO-RE taxonomy. *Psychology & Health 26*(11), 1479–1498.

Mihailidis, A., J. N. Boger, T. Craig, and J. Hoey (2008). The COACH prompting system to assist older adults with dementia through handwashing: An efficacy study. *BMC Geriatrics 8*(1), 1–18.

Mikolajczak, J., O. A. Blanson Henkemans, and J. F. E. M. Keijsers (2011). Ehealth Analyse en Sturingsinstrument (eASI): Ontwikkeling en Toepassing Versie 1.0. *Tijdschrift voor gezondheidswetenschappen 89*(2), 78–82.

Miller, A. L., M. A. Horodynski, H. E. B. Herb, K. E. Peterson, D. Contreras, N. Kaciroti, J. Staples-Watson, and J. C. Lumeng (2012). Enhancing self-regulation as a strategy for obesity prevention in Head Start preschoolers: The growing healthy study. *BMC public health 12*(1), 1040.

Milyavskaya, M. and M. Inzlicht (2017). Attentional and Motivational Mechanisms of Self-Control. In *Routledge International Handbook of Self-Control in Health and Well-Being*, pp. 11–23. Routledge.

Miserandino, M. (1996). Children Who Do Well in School: Individual Differences in Perceived Competence and Autonomy in Above-Average Children. *Journal of educational psychology 88*(2), 203.

Moran, M. (2003). *Rethinking the Reasonable Person: An Egalitarian Reconstruction of the Objective Standard*. Oxford University Press on Demand.

Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics 31*, 175–178.

Moray, N. and T. Inagaki (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science 1*, 354–365.

Moray, N., T. Inagaki, and M. Itoh (2000). Adaptive Automation, Trust, and Self-Confidence in Fault Management of Time-Critical Tasks. *Journal of Experimental Psychology: Applied 6*(1), 44.

Mosier, K. L., M. Dunbar, L. McDonnell, L. J. Skitka, M. Burdick, and B. Rosenblatt (1998). Automation Bias and Errors: Are Teams Better than Individuals? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Volume 42, pp. 201–205.

Mosier, K. L. and L. J. Skitka (1999). Automation Use and Automation Bias. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Volume 43, pp. 344–348.

Mosier, K. L., L. J. Skitka, S. Heers, and M. Burdick (1998). Automation Bias: Decision Making and Performance in High-Tech Cockpits. *The International Journal of Aviation Psychology 8*, 47–63.

Mullainathan, S. and E. Shafir (2013). *Scarcity: Why Having Too Little Means So Much*. New York, NY: Time Books, Henry Holt & Company LLC.

Muraven, M. and R. F. Baumeister (2000). Self-Regulation and Depletion of Limited Resources: Does Self-Control Resemble a Muscle? *Psychological bulletin 126*(2), 247.

Muraven, M. and E. Slessareva (2003). Mechanisms of Self-Control Failure: Motivation and Limited Resources. *Personality and Social Psychology Bulletin 29*(7), 894–906.

Narayanan, A. and V. Shmatikov (2008). Robust De-Anonymization of Large Sparse Datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125. IEEE.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of general psychology 2*(2), 175.

Nissenbaum, H. (2011). A Contextual Approach to Privacy Online. *Daedalus 140*(4), 32–48.

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

Noggle, R. (1996). Manipulative Actions: A Conceptual and Moral Analysis. *American Philosophical Quarterly 33*(1), 43–55.

Nozick, R. (1969). Coercion. In S. Morgenbesser, P. Suppes, and M. White (Eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, pp. 440–472. New York, NY: St. Martin's Press.

Obermair, C., W. Reitberger, A. Meschtscherjakov, M. Lankes, and M. Tscheligi (2008). perFrames: Persuasive Picture Frames for Proper Posture. In H. Oinas-Kukkonen, P. Hasle, M. Harjumaa, K. Segerståhl, and P. Øhrstrøm (Eds.), *Persuasive Technology*, Volume 5033 of *Lecture Notes in Computer Science*,

pp. 128–139. Springer Berlin Heidelberg.

Ochoa, S. F. and F. J. Gutierrez (2018). Architecting E-Coaching Systems: A First Step for Dealing with Their Intrinsic Design Complexity. *Computer 51*(3), 16–23.

Oettingen, G., D. Mayer, and J. Thorpe (2010). Self-regulation of commitment to reduce cigarette consumption: Mental contrasting of future with reality. *Psychology and Health 25*(8), 961–977.

Oinas-Kukkonen, H. (2010). Behavior Change Support Systems: A Research Model and Agenda. In T. Ploug, P. Hasle, and H. Oinas-Kukkonen (Eds.), *Persuasive Technology*, Volume 6137 of *Lecture Notes in Computer Science*, pp. 4–14. Springer Berlin Heidelberg.

Oinas-Kukkonen, H. (2013). A foundation for the study of behavior change support systems. *Personal and Ubiquitous Computing 17*(6), 1223–1235.

Orbell, S., S. Hodgkins, and P. Sheeran (1997). Implementation Intentions and the Theory of Planned Behavior. *Personality and Social Psychological Bulletin* (23), 945–954.

O'Regan, K. and A. Noë (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences 23*, 939–973.

Oshana, M. A. L. (1998). Personal Autonomy and Society. *Journal of Social Philosophy 29*(1), 81–102.

Oshana, M. A. L. (2006). *Personal Autonomy in Society*. Burlington, VT: Ashgate.

Parasuraman, R. and D. H. Manzey (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society 52*(3), 381–410.

Parasuraman, R., R. Molloy, and I. L. Singh (1993). Performance Consequences of Automation-Induced 'Complacency'. *The International Journal of Aviation Psychology 3*(1), 1–23.

Peacocke, C. (1979). Deviant Causal Chains. *Midwest Studies In Philosophy 4*(1), 123–155.

Philips (2010). DirectLife. `http://www.philips.nl/a-w/about/news/archive/standard/about/news/press/20100802_directlife.html`. Retrieved January 4th 2017.

Power, D. J. (2008). Decision Support Systems: A Historical Overview. In *Handbook on Decision Support Systems 1*, International Handbooks Information System, pp. 121–140. Springer Berlin Heidelberg.

Procee, R., B. A. Kamphorst, A. Van Wissen, and J.-J. Meyer (2014). An Agent-Based Model of Procrastination. In *ECAI 2014*, Volume 263 of *Frontiers in Artificial Intelligence and Applications*, pp. 747–752.

Rao, A. S. and M. P. Georgeff (1995). BDI Agents: From Theory to Practice. In *ICMAS*, Volume 95, pp. 312–319.

Rashidi, P., D. J. Cook, L. B. Holder, and M. Schmitter-Edgecombe (2011). Discovering Activities to Recognize and Track in a Smart Environment.

IEEE transactions on knowledge and data engineering 23(4), 527–539.

Raz, J. (1986). *The Morality of Freedom*. Oxford, UK: Clarendon.

Regan, P. M. (2000). *Legislating Privacy: Technology, Social Values, and Public Policy*. University of North Carolina Press.

Reichenbach, J., L. Onnasch, and D. Manzey (2010). Misuse of automation: The impact of system experience on complacency and automation bias in interaction with automated aids. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Volume 54, pp. 374–378.

Reisch, M. (2002). Defining Social Justice in a Socially Unjust World. *Families in Society 83*(4), 343–354.

Richardson, H. S. (1997). *Practical Reasoning about Final Ends*. Cambridge University Press.

Roessler, B. (2008). New Ways of Thinking about Privacy. In J. S. Dryzek, B. Honig, and A. Phillips (Eds.), *The Oxford Handbook of Political Theory*. Oxford University Press.

Roessler, B. and D. Mokrosinska (2013). Privacy and social interaction. *Philosophy & Social Criticism 39*(8), 771–791.

Rowlands, M. (2009). The Extended Mind. *Zygon® 44*(3), 628–641.

Rupert, R. D. (2004). Challenges to the Hypothesis of Extended Cognition. *The Journal of Philosophy 101*(8), 389–428.

Russell, S. J. and P. Norvig (2003). *Artificial Intelligence: A Modern Approach*. New Jersey: Prentice Hall. Second Edition.

Samuelson, W. and R. Zeckhauser (1988). Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty 1*(1), 7–59.

Schmeichel, B. J., K. D. Vohs, and R. F. Baumeister (2003). Intellectual Performance and Ego Depletion: Role of the Self in Logical Reasoning and Other Information Processing. *Journal of Personality and Social Psychology 85*(1), 33–46.

Schmeichel, B. J. and A. Zell (2007). Trait Self-Control Predicts Performance on Behavioral Tests of Self-Control. *Journal of Personality 75*(4), 743–756.

Schmidt, D. P. (2007). Reasonable Person Standard. In R. W. Kolb (Ed.), *Encyclopedia of Business Ethics and Society*. Thousand Oaks: Sage Publications.

Schutz, P. A. and H. A. Davis (2000). Emotions and Self-Regulation During Test Taking. *Educational psychologist 35*(4), 243–256.

Sheeran, P. and C. Abraham (2003). Mediator of Moderators: Temporal Stability of Intention and the Intention-Behavior Relation. *Personality and Social Psychology Bulletin 29*(2), 205–215.

Sheeran, P. and S. Orbell (1999). Implementation intentions and repeated behaviour: Augmenting the predictive validity of the theory of planned behaviour. *European Journal of Social Psychology* (29), 349–369.

Sheeran, P. and S. Orbell (2000). Using Implementation Intentions to Increase Attendance for Cervical Cancer Screening. *Health Psychology 18*, 283–289.

Sheeran, P., T. L. Webb, and P. M. Gollwitzer (2005). The Interplay Between Goal Intentions and Implementation Intentions. *Personality and Social Psychology Bulletin 31*(1), 87–98.

Sinhababu, N. (2013). The Desire–Belief Account of Intention Explains Everything. *Noûs 47*(4), 680–696.

Sirois, F. M. (2004). Procrastination and intentions to perform health behaviors: The role of self-efficacy and the consideration of future consequences. *Personality and Individual Differences 37*, 115–128.

Sirois, F. M., M. L. Melia-Gordon, and T. A. Pychyl (2003). "I'll look after my health, later": an investigation of procrastination and health. *Personality and Individual Differences 35*(5), 1167–1184.

Skitka, L. J., K. L. Mosier, and M. Burdick (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies 51*(5), 991–1006.

Slors, M. (2015). Conscious intending as self-programming. *Philosophical Psychology 28*(1), 94–113.

Smids, J. (2012). The Voluntariness of Persuasive Technology. In M. Bang and E. L. Ragnemalm (Eds.), *Persuasive Technology. Design for Health and Safety*, Volume 7284 of *Lecture Notes in Computer Science*, pp. 123–132. Springer Berlin Heidelberg.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical studies 132*(1), 99–107.

Sterelny, K. (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences 9*(4), 465–481.

Stolk, M. (2015). MOOD SUPPORT: An Adaptive Social-Context Aware Mood Support System. Master's thesis, VU University, Amsterdam, the Netherlands.

Strine, T. W. and D. P. Chapman (2005). Associations of frequent sleep insufficiency with health-related quality of life and health behaviors. *Sleep Medicine 6*(1), 23–27.

Stroud, S. (2010, Jan). Is Procrastination Weakness of Will? In *The Thief of Time: Philosophical Essays on Procrastination*, pp. 51–67.

Stunkel, L., M. Benson, L. McLellan, N. Sinaii, G. Bedarida, E. Emanuel, and C. Grady (2010). Comprehension and Informed Consent: Assessing the Effect of a Short Consent Form. *IRB 32*(4), 1.

Sunstein, C. R. (2015). Fifty Shades of Manipulation. *Journal of Marketing Behavior 1*, 213–244.

Susser, D., B. Roessler, and H. Nissenbaum (2018). Online Manipulation: Hidden Influences in a Digital World. *Georgetown Law Technology Review*. Forthcoming.

Susser, D., B. Roessler, and H. Nissenbaum (2019). Technology, autonomy, and manipulation. *Internet Policy Review 8*(2).

Sutton, J. (2010). Exograms and Interdisciplinarity: History, the Extended

Mind and the Civilizing Process. In R. Menary (Ed.), *The Extended Mind*, pp. 189–225. Cambridge, MA: MIT Press.

Sutton, J., C. B. Harris, P. G. Keil, and A. J. Barnier (2010). The psychology of memory, extended cognition, and socially distributed remembering. *Phenomenology and the Cognitive Sciences 9*(4), 521–560.

Tangney, J. P., R. F. Baumeister, and A. L. Boone (2004). High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success. *Journal of Personality 72*(2), 271–324.

Taylor, S. E., L. B. Pham, I. D. Rivkin, and D. A. Armor (1998). Harnessing the Imagination: Mental Simulation, Self-Regulation, and Coping. *American Psychologist 53*(4), 429.

Thackray, R. I. and R. M. Touchstone (1989). Detection Efficiency on an Air Traffic Control Monitoring Task with and without Computer Aiding. *Aviation, Space, and Environmental Medicine 60*, 744–748.

Thaler, R. H. and C. R. Sunstein (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. London, UK: Penguin Books.

Thayer, R. E., J. R. Newman, and T. M. McClain (1994). Self-Regulation of Mood: Strategies for Changing a Bad Mood, Raising Energy, and Reducing Tension. *Journal of personality and social psychology 67*(5), 910.

Thompson, R. F. (2009). Habituation: A History. *Neurobiology of learning and memory 92*(2), 127.

Thøgersen, J. (2003). Monetary Incentives and Recycling: Behavioural and Psychological Reactions to a Performance-Dependent Garbage Fee. *Journal of Consumer Policy 26*(2), 197–228.

Tice, D. M., E. Bratslavsky, and R. F. Baumeister (2001). Emotional Distress Regulation Takes Precedence Over Impulse Control: If You Feel Bad, Do It! *Journal of Personality and Social Psychology 80*(1), 53–67.

Torma, G., J. Aschemann-Witzel, and J. Thøgersen (2018). I nudge myself: Exploring 'self-nudging' strategies to drive sustainable consumption behaviour. *International Journal of Consumer Studies 42*(1), 141–154.

Tversky, A. and D. Kahneman (1991). Loss Aversion in Riskless Choice: A Reference Dependent Model. *Quarterly Journal of Economics 106*, 1039–1061.

Unwin, N. (1985). Relativism and Moral Complacency. *Philosophy 60*(232), 205–214.

Van Baal, P. H. M., R. Heijink, R. T. Hoogenveen, and J. J. Polder (2007). Zorgkosten van ongezond gedrag. Zorg voor euro's - 3. Last visited in December 2019.

Van de Ven, P., M. R. Henriques, M. Hoogendoorn, M. Klein, E. McGovern, J. Nelson, H. Silva, and E. Tousset (2012). A Mobile System for Treatment of Depression. In P. Cipresso, M. Hoogendoorn, M. Klein, and A. Matic (Eds.), *Computing Paradigms for Mental Health. Proceedings of MindCare 2012, in conjunction with BIOSTEC 2012*, pp. 47–58.

Van Wissen, A. (2014). *Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains*. Ph. D. thesis, VU University Amsterdam.

Van Wissen, A., B. A. Kamphorst, and R. Van Eijk (2013). A Constraint-Based Approach to Context. In P. Brézillon, P. Blackburn, and R. Dapoigny (Eds.), *Proceedings of the Eighth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT'13)*, pp. 171–184. Springer.

Varela, F. J., E. Thompson, and E. Rosch (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Velleman, D. (1992). What Happens When Someone Acts? *Mind 101*, 461–481.

Velleman, J. D. (2000). On the Aim of Belief. In *The Possibility of Practical Reason*, pp. 244–281. Oxford, UK: Oxford University Press.

Vohs, K. D. and R. F. Baumeister (2016). *Handbook of self-regulation: Research, theory, and applications*. Guilford Publications.

Vohs, K. D., R. F. Baumeister, B. J. Schmeichel, J. M. Twenge, N. M. Nelson, and D. M. Tice (2008). Making Choices Impairs Subsequent Self-Control: A Limited-Resource Account of Decision Making, Self-Regulation, and Active Initiative. pp. 45–77.

Wallace, R. J. (2001). Normativity, Commitment, and Instrumental Reason. *Philosophers' Imprint 1*(3), 1–26.

Warner, T. (2012). E-coaching systems: Convenient, anytime, anywhere, and nonhuman. *Performance Improvement 51*(9), 22–28.

Webb, T. L. and P. Sheeran (2007). How do implementation intentions promote goal attainment? A test of component processes. *Journal of Experimental Social Psychology 43*(2), 295–302.

Weizenbaum, J. (1966). ELIZA—a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM 9*(1), 36–45.

Wertheimer, A. (1987). *Coercion*. Princeton, NJ: Princeton University Press.

Westlund, A. C. (2009). Rethinking Relational Autonomy. *Hypatia 24*(4), 26–49.

Wheeler, M. (2010). In Defense of Extended Functionalism. In R. Menary (Ed.), *The Extended Mind*, pp. 245–270. Cambridge, MA: MIT Press.

Wheeler, M. (2011). Embodied Cognition and the Extended Mind. *The continuum companion to philosophy of mind*, 220–238.

Wickens, C. D., B. A. Clegg, A. Z. Vieane, and A. L. Sebok (2015). Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society 57*(5), 728–739.

Wilson, R. and A. Clark (2009). How to situate cognition: Letting nature take its course. In P. Robbins and M. Aydede (Eds.), *The Cambridge handbook of situated cognition*, pp. 55–77. New York, NY: Cambridge University Press.

Wilson, R. A. (2004). *Boundaries of the Mind: The Individual in the Fragile Sciences - Cognition*. Cambridge, UK: Cambridge University Press.

Wood, R. and A. Bandura (1989). Impact of Conceptions of Ability on Self-Regulatory Mechanisms and Complex Decision Making. *Journal of personality and social psychology 56*(3), 407.

Wooldridge, M. and N. R. Jennings (1995). Agent Theories, Architectures, and Languages: A Survey. In M. J. Wooldridge and N. . R. Jennings (Eds.), *Intelligent Agents*, Volume 890 of *Lecture Notes in Computer Science*, pp. 1–39. Springer Berlin Heidelberg.

Xiao, Q., H. Arem, S. C. Moore, A. R. Hollenbeck, and C. E. Matthews (2013). A Large Prospective Investigation of Sleep Duration, Weight Change, and Obesity in the NIH-AARP Diet and Health Study Cohort. *American Journal of Epidemiology 178*(11), 1600–1610.

Yeung, K. (2017). 'Hypernudge': Big Data as a Mode of Regulation by Design. *Information, Communication & Society 20*(1), 118–136.

Zimmerman, B. J. (2000). Attaining Self-Regulation: A Social Cognitive Perspective. In *Handbook of self-regulation*, pp. 13–39. Elsevier.

Zimmerman, B. J. (2008). Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. *American educational research journal 45*(1), 166–183.

Zimmerman, B. J. and A. Kitsantas (1997). Developmental Phases in Self-Regulation: Shifting From Process Goals to Outcome Goals. *Journal of educational psychology 89*(1), 29.

Samenvatting in het Nederlands

Deze dissertatie beoogt een bijdrage te leveren aan het verantwoord innoveren van geautomatiseerde systemen die individuen dynamische, gepersonaliseerde en persuasieve ondersteuning bieden bij (aspecten van) hun vermogen om hun eigen gedrag en emoties te reguleren (*zelfregulatie*). Deze bijdrage bestaat onder andere uit het bieden van conceptuele opheldering, het schetsen van het ethische landschap omtrent de ontwikkeling van deze technologieën, en het leveren van praktische richtlijnen voor het ontwikkelen, gebruiken, en reguleren van deze technologieën.

In **Hoofdstuk 1** plaats ik het begrip 'zelfregulatiefalen' in verhouding tot de filosofisch beter bekende termen 'akrasia' en 'wilszwakte'. Ik concludeer hierbij dat 'zelfregulatiefalen' een breder begrip is dat meer gevallen van menselijk falen omvat en tegelijk meer complexiteit openbaart van de causale mechanismes die ten grondslag liggen aan verschillende vormen van menselijk falen. Vervolgens zet ik uiteen hoe ik de termen zelfregulatie en zelfregulatiefalen begrijp en beargumenteer ik dat beide begrippen nauwer begrepen moeten worden dan hoe ze soms in de literatuur gebruikt worden. Ik bespreek daarna mogelijke interventies die ontwikkeld kunnen worden ter ondersteuning van zelfregulatieprocessen en ik focus daarbij met name op de rol die technologie kan spelen in het bieden van een coachende vorm van ondersteuning die vierentwintig uur per dag en zeven dagen in de week beschikbaar is, dynamisch afgestemd wordt op het gedrag en de specifieke voorkeuren van het individu, en persuasieve technieken gebruikt om effectief advies te bieden ('e-coaching'). Vervolgens bespreek ik drie perspectieven vanwaaruit er argumenten (kunnen) worden gegeven ten gunste van de ontwikkeling van technologische interventies die ten doel hebben mensen te ondersteunen bij hun zelfregulatie. Deze perspectieven zijn respectievelijk het perspectief van de sociale impact (bijvoorbeeld "zelfregulatiefalen van individuen met betrekking tot alcoholgebruik leidt tot grote kosten in de gezondheidszorg"), het perspectief van gezondheid en welzijn van het individu

(bijvoorbeeld "zelfregulatiefalen van individuen met betrekking tot slaap leidt tot verminderde concentratie bij het betreffende individu"), en het perspectief van zelfbeschikking ('self-governing agency'; bijvoorbeeld "zelfregulatiefalen ondermijnt effectief de sturing die een individu aan zijn of haar leven wil geven"). Ik besluit het hoofdstuk door te benadrukken dat de redenen ten gunste van de ontwikkeling van technologische interventies uit elk van deze perspectieven afgewogen dienen te worden tegen mogelijk zwaarwegendere redenen ten nadele van deze technologieën die voortkomen uit ethische bezwaren in relatie tot het wijdverspreid gebruik van deze vorm van technologische ondersteuning van zelfregulatie.

In **Hoofdstuk 2** bespreek ik vervolgens bestaande definities van 'e-coachingsystemen' om daarna te beargumenteren dat die definities te veelomvattend zijn. Ik bied zelf een nauwere definitie als alternatief, waarin expliciet wordt gemaakt dat e-coachingsystemen zich onderscheiden van andere vormen van zelfregulatieondersteuning door de mate van geavanceerdheid en onafhankelijkheid die dit type systeem karakteriseert. Door deze scherpere definitie komt beter in beeld welke specifieke problemen kunnen ontstaan als gevolg van het gebruik van geavanceerde ondersteuningstechnologieën. In de laatste sectie van het hoofdstuk bespreek ik de relatie tussen gebruiker en e-coachingsysteem in het licht van recente literatuur over 'extended will' en 'extended mind', de respectievelijke theses dat cognitie en wilskracht zich kunnen uitstrekken voorbij de grenzen van het menselijk lichaam en (tijdelijk) delen van de omgeving kunnen bevatten als deel van een cognitief of volitioneel systeem. Vooruitlopend op het argument aangaande *zelfvoldoening* ('complacency') dat nader uitgewerkt wordt in Hoofdstuk 4, beargumenteer ik dat het wellicht niet wenselijk is om het soort nauwe integratie tussen gebruiker en e-coachingsysteem na te streven dat noodzakelijk is om te kunnen spreken van een samengesteld systeem, wanneer dat zou betekenen dat gebruikers afstand zouden doen van bepaalde verantwoordelijkheden die gerelateerd zijn aan het zelf bepalen van de koers die ze willen voeren in hun eigen leven.

In **Hoofdstuk 3** bied ik een initiële inventarisatie van ethische zorgen die opkomen in relatie tot het wijdverspreid gebruik van e-coachingsystemen. Deze inventarisatie is bedoeld als leidraad voor toekomstige discussies over e-coachingsystemen maar functioneert tevens als kader waarbinnen ik het vraagstuk van zelfvoldoening positioneer. Ik onderscheid in de inventarisatie drie categorieën van ethische zorgen. De eerste categorie betreft onderwerpen die betrekking hebben op sociale rechtvaardigheid, zoals gelijke toegang (krijgt iedereen dezelfde kansen om dezelfde kwaliteit ondersteuning te krijgen?), vrijheid (blijven we vrij om deze technologieën te weren?), en welzijn (hoe kijken we vanuit een sociaal-cultureel perspectief aan tegen mensen wier zelfregulatie ondersteund wordt door e-coachingsystemen?). De tweede categorie van ethische zorgen die ik bespreek betreft onderwerpen die betrekking

hebben op het maken van inbreuk op bepaalde rechten. Specifiek benoem ik in deze context de prangende kwestie van privacy (hoe dienen we om te gaan met de expliciete conclusies die een systeem trekt uit verzamelde data?) en de risico's op vormen van dwang en manipulatie (hoe weren we ons tegen het risico dat e-coaching gericht kan zijn op de verkoop van bepaalde gezondheidsproducten in plaats de bevordering van de gezondheid van het individu?). De derde en laatste categorie die ik bespreek in het hoofdstuk betreft zorgen die betrekking hebben op onbedoelde gevolgen van de specifieke wisselwerking tussen gebruiker en e-coachingsysteem. In het bijzonder richt ik me hierbij op de mogelijke negatieve effecten op het uitoefenen van zelfsturing ('self-governance') door de facilitatie van zelfvoldoening in relatie tot praktisch redeneren. Het centrale idee hier is dat er een keerzijde is aan de persuasieve technieken die e-coachingsystemen effectief zouden moeten maken, namelijk dat ze gebruikers kunnen bewegen tot het maken van de onjuiste inschatting dat ze, enkel door het volgen van advies van een door henzelf ingeschakeld e-coachingsysteem, voldoende inspanning hebben geleverd aan het zelf bepalen van de koers in hun leven. Ik besluit het hoofdstuk met een eerste verkenning van deze zorg.

In **Hoofdstuk 4** staat de nadere uitwerking van het vraagstuk over zelfvoldoening centraal. In voorbereiding op een analyse van dit vraagstuk introduceer ik eerst een reeks relevante begrippen uit de handelingstheorie van Michael Bratman. Een belangrijk achterliggend idee van deze theorie is dat menselijk handelen inherent temporeel is en dat het vormen van intenties en het maken van plannen behoren tot de kernmechanismes voor mensen om handelingen over tijd te coördineren en zo sturing te geven aan hun leven. Met het belang van intenties en plannen voor ogen krijgen we ook beter in beeld wat het risico is voor onze zelfsturing wanneer we een te weinig kritische houding aannemen jegens externe invloeden op de processen van het vormen van intenties en het maken van plannen. Immers, zo betoog ik later in het hoofdstuk, handelen op advies dat onzorgvuldig is overwogen door een individu in relatie tot zijn of haar eigen doelen, plannen, en waardes, dwarsboomt zowel zelfsturing in het moment ('synchronic self-governance') als zelfsturing over tijd ('diachronic self-governance').

Met het begrippenapparaat vastgesteld, introduceer en bespreek ik achtereenvolgens twee claims die betrekking hebben tot het vraagstuk van zelfvoldoening. De eerste claim is van empirische aard en stelt dat e-coachingsystemen, door de manier waarop ze functioneren, het risico verhogen op het ontstaan van zelfvoldoening in relatie tot praktisch redeneren ('Heightened Risk of Complacency (HRC) Claim'). Om deze claim kracht bij te zetten bespreek ik empirische literatuur uit het domein van beslissingsondersteunende systemen ('decision support systems'; DSS) die aantoont dat mensen die in hun besluitvorming ondersteund worden door geautomatiseerde hulpsystemen

ontvankelijk zijn voor het maken van bepaalde typen fouten. Hierbij gaat het om het negeren van relevante, waarneembare probleemsignalen uit andere bronnen omdat het geautomatiseerde hulpsysteem deze signalen abusievelijk niet onderschrijft ('errors of omission'), of juist het handelen op basis van onjuist advies van het geautomatiseerde hulpsysteem ('errors of commission'). De suggestie achter de HRC Claim is dat het gebruik van e-coachingsystemen eveneens het risico op het begaan van dit type fouten zal vergroten.

De tweede claim is een conceptuele claim die uiting geeft aan het idee dat zelfvoldoening ten aanzien van praktisch redeneren negatieve gevolgen heeft voor de mate van zelfsturing die we kunnen toeschrijven aan een individu ('Implication for Self-Governance (ISG) Claim'). Deze claim kent twee vooronderstellingen. De eerste vooronderstelling is dat zelfsturing een praktisch standpunt van het individu vereist waaruit de handeling volgt (zelfsturing in het moment) of de handelingen volgen (zelfsturing over tijd). Een praktisch standpunt in deze context kan gezien worden als een coherente en consistente constellatie van overtuigingen, intenties, en plannen. De tweede vooronderstelling, die volgt uit de eerste, is dat zelfsturing een bepaalde mate van waakzaamheid ('vigilance') vereist ten aanzien van processen omtrent het vormen van een praktisch standpunt. In andere woorden, individuen dienen een actieve houding te hebben in het zich verhouden tot zowel interne invloeden (zoals verlangens of wensen) en externe invloeden (zoals suggesties van menselijke coaches of e-coachingsystemen). Uit de ISG Claim volgt dat er sprake is van verminderde zelfsturing wanneer er bij een individu zelfvoldoening ontstaat in het praktisch redeneren, zodanig dat het individu tekort schiet in het zich adequaat verhouden tot dergelijke invloeden.

De 'Heightened Risk of Complacency Claim' en de 'Implication for Self-Governance Claim' vormen tezamen het vraagstuk over zelfvoldoening. Het hoofdstuk besluit met een reeks fictieve voorbeelden die beogen de risico's van zelfvoldoening te illustreren en het belang van het vraagstuk te onderstrepen.

In **Hoofdstuk 5** bespreek ik de implicaties die volgen als zelfvoldoening inderdaad het probleem zal blijken te zijn dat ik geschetst heb. Daarnaast geef ik concrete aanbevelingen om zelfvoldoening tegen te gaan die gericht zijn op gebruikers en ontwikkelaars van e-coachingsystemen, alsmede op beleidsmakers op het gebied van deze technologieën. De belangrijkste implicatie voor gebruikers van e-coachingsystemen is dat zij een actieve houding dienen in te nemen ten aanzien het beoordelen van externe suggesties die ze niet zelf hebben bedacht en ook nog niet eerder hebben gezien. Ongeacht hoe vaak een systeem in het verleden goede suggesties heeft gedaan, dient een gebruiker waakzaam te blijven en na te gaan of een gedane suggestie passend is ten aanzien van het beoogde doel, alsmede of de suggestie consistent is binnen het grotere geheel van zijn of haar doelen, waardes en plannen.

Designers van e-coachingsystemen kunnen dit proces van individuele ge-

bruikers ondersteunen door de dialoog tussen gebruiker en systeem zo te ontwerpen dat het individu actief betrokken blijft bij het vormgeven van het coachingsproces. Dit kan bijvoorbeeld door suggesties op te volgen met evaluatievragen over de suggesties. Daarnaast kunnen designers ook expliciet rekening houden met het risico op zelfvoldoening door technieken te ontwikkelen die zelfvoldoening tegengaan, zoals het aanbieden van informatie over hoe een e-coachingsysteem tot een bepaalde suggestie is gekomen en hoeveel vertrouwen het systeem heeft in de suggestie.

Een belangrijke implicatie voor beleidsmakers is dat een terughoudende attitude op zijn plaats is ten aanzien van aanmoedigingsprogramma's en verplichtstelling van het gebruik van e-coachingsystemen. In dit kader pleit ik voor het oprichten van een onafhankelijk instituut dat over expertise beschikt om e-coachingsystemen te toetsen aan bepaalde standaarden—waaronder een nog te ontwikkelen standaard aangaande de mate waarin een e-coachingsysteem zelfvoldoening faciliteert—en dat een kwaliteitscertificering kan uitdelen wanneer een systeem aan de standaarden voldoet. Een dergelijk instituut zal mijns inziens zowel individuen ondersteunen bij het maken van een goede inschatting over welke e-coachingsystemen geschikt zijn voor hen, als beleidsmakers helpen om geïnformeerde beslissingen te nemen in domeinen waarin regulatie van e-coachingsystemen wenselijk is.

In het laatste gedeelte van het hoofstuk doe ik vijf concrete aanbevelingen aan designers en ontwikkelaars van e-coachingsystemen voor het tegengaan van zelfvoldoening. *De eerste aanbeveling* heeft betrekking tot het verkrijgen van toestemming ('informed consent') van de gebruiker om data te verzamelen, te verwerken, en te gebruiken om persuasieve suggesties te doen. Gegeven het adaptieve karakter van e-coachingsystemen, is het voor gebruikers moeilijk van tevoren te overzien hoe e-coachingsystemen zich zullen aanpassen aan veranderende omstandigheden en gedragingen over tijd. Gezien dat gegeven is de aanbeveling een coherent pakket van informatie, begripstoetsen, en instemmingsvragen te ontwikkelen en die gaandeweg het coachingsproces in te zetten om zo gerichte instemming te verkrijgen voor het gebruik van technieken die op dat moment voor de gebruiker relevant zijn binnen het coachingsproces. *De tweede aanbeveling* is gebruikers beter in staat te stellen de validiteit en geschiktheid van een suggestie te toetsen door de gebruiker inzicht te verschaffen in de redenering achter een suggestie. Deze aanbeveling vindt aansluiting bij recente ontwikkelingen op het onderzoeksgebied van 'Explainable A.I.'. *De derde aanbeveling* is gebruikers beter bewust te maken van de faalbaarheid van het systeem. Dit heeft ten doel overschatting van de betrouwbaarheid van het systeem tegen te gaan. *De vierde aanbeveling* is gebruikers expliciet gelegenheden te bieden om doelen te herzien en tevens het gebruik van het systeem zelf te herevalueren. Het doel hiervan is de waakzaamheid van de gebruiker te ondersteunen. *De vijfde en laatste aanbeveling* is

e-coachingsystemen expliciet aan gebruikers te laten vragen om een suggestie kritisch te beschouwen op het moment dat deze aangeboden wordt, en tevens evaluatievragen te laten stellen aan gebruikers over gedane suggesties in het verleden. Hoewel dergelijke vragen spaarzaam ingezet dienen te worden om de dialoog tussen gebruiker en systeem niet onnodig te hinderen, kunnen welgeplaatste vragen bijdragen aan een vorm van wederzijdse communicatie waarin de gebruiker actief deelneemt aan de vormgeving van het coachingsproces.

In de **Conclusie**, ten slotte, vat ik de belangrijkste punten uit de dissertatie samen en benoem ik vier contributies die deze dissertatie mijns inziens maakt. *De eerste contributie* bestaat uit de conceptuele opheldering die ik heb geboden binnen verscheidene literaturen, bijvoorbeeld door het opstellen van een nieuwe definitie van e-coachingsystemen. *De tweede contributie* bestaat uit het in kaart brengen van het ethische landschap omtrent e-coachingsystemen. *De derde contributie* bestaat uit de filosofische analyse van de manier waarop e-coachingsystemen een vorm van zelfvoldoening in de hand kunnen werken die negatieve gevolgen heeft voor zelfsturing. *De vierde contributie* bestaat uit de set van aanbevelingen die ik heb geboden om zelfvoldoening tegen te gaan. De hoop is dat deze contributies tezamen de weg vrijmaken voor verdere verantwoordelijke innovatie op het gebied van geautomatiseerde ondersteuning van zelfregulatie.

Acknowledgements

In the years working towards the completion of this dissertation I have been fortunate enough to have been supported, encouraged and inspired by a great many individuals. It is my pleasure to express my gratitude here to all who have graciously invested their time, offered advice, or lent a listening ear.

To begin with, I am deeply indebted to Joel Anderson, who has been unwavering in his support from the moment we first started working together. Always engaging, always maintaining a thoughtful balance between constructive criticism and uplifting words of encouragements, always ready to make connections and introductions, always mindful of my well-being, Joel has gone above and beyond in his role as promotor. I would also like to thank Marcus Düwell for his insightful comments on various chapter drafts and for his support when it mattered.

I wish to extend my thanks to the members of my reading committee—Annemarie Kalis, Anthonie Meijers, Ingrid Robeyns, Peter-Paul Verbeek, and Paul Ziche—for taking the time to critically engage with my work and for not being complacent in this regard. A special thanks goes out to Annemarie for providing in-depth comments on multiple drafts of one of the chapters. Thanks also to Jan Broersen, Sven Nyholm, and Denise de Ridder for graciously agreeing to be on my defense committee. I am also very appreciative of the Department of Philosophy and Religious Studies as a whole for the many years of providing a demanding but supportive working environment. Thanks in particular to Niels van Miltenburg, Jesse Mulder, Sem de Maagt and (former Utrecht colleague) Stephen Riley, as well as Suzanne van Vliet, Judith Zijm, and Biene Meijerman. I have also benefited from feedback from the members of the Practical Philosophy (PF) Colloquium.

As my PhD was an interdisciplinary endeavor, I was privileged to collaborate with colleagues from other Utrecht University departments as well. At the Department of Information and Computing Sciences I wish to extend my gratitude to John-Jules Meyer; at Utrecht's Self-Regulation Lab I wish to thank again Denise de Ridder, as well as Floor Kroese, Marieke Adriaanse, Catharine Evers, and former lab member Sanne Nauts. A special word of thanks goes out

to Virginia Dignum, who was at Utrecht University when I started my PhD, for taking me under her wing in the early stages of my academic career.

Over the years, I have also had the pleasure to meet and interact with many terrific researchers outside of Utrecht who, each in their own way, have helped shape my thinking on various subjects. At one point or another, I have benefited from conversations with Roy Baumeister, Peter Gollwitzer, Shaun Gallagher, John Lysaker, Neil Levy, and Alfred Mele. At Philips Research, my thanks goes out to Saskia van Dantzig, Reinder Haakma, and Aart van Halteren. Thanks also to Michel Klein, Julienka Mollee, and all the other researchers participating in the Healthy Lifestyle Solutions programme for creating an atmosphere and culture conducive to interdisciplinary collaboration.

I owe sincere gratitude to Stanford's Philosophy Department for hosting me from April–June 2014 and to Michael Bratman for generously agreeing to be my sponsor. To say that I learnt a great deal from Michael about doing philosophy right is a major understatement. While at Stanford, I had the pleasure of meeting Jens Gillessen, who has kindly provided comments on one of the chapter drafts. I am also thankful for enlightening conversations with Tamar Schapiro.

In the fall of 2014 I was welcomed at Carlton University's Department of Psychology for an extended stay. There, again, I met a number of wonderful people, among whom Shamarukh Chowdhury, Mohsen Haghbin, and Eve-Marie Blouin-Hudon. I feel enormous gratitude towards Timothy Pychyl and his wife Beth Rohr for opening their home to me and for the friendship that ensued.

I am grateful to Mark Hoogendoorn, Guszti Eiben, Aart van Halteren, and Evert Haasdijk for inviting me to join their team at the Computational Intelligence Research Group of the Vrije Universiteit Amsterdam. Thanks to the members of that group, especially Ali El Hassouni, Eoin Grua, and Alessandro Zonta, for the great atmosphere during my time at the VU. I am indebted to Guszti, Mark, Dick Bulterman, and the Department of Computer Science, for providing me with time to finish this book.

I have been privileged to have been surrounded by a wonderful group of friends during the writing of this dissertation. Thanks to Charlotte Gerritsen, Tibor Bosse, Fiemke Griffioen-Both, Mark Hoogendoorn, Rianne van Lambalgen, and Arul Elangovan for the many fun dinners and outings over the years. A number of friendships date back even longer, and I am incredibly grateful for those continued friendships. While I cannot name everyone here, I do wish to mention in particular Inge Wolsky–Wiersma, Joke van Haren–Hogeling, Kristyanne Flos, Lotte Kramer, Melanie Philippi–Kroes, Maarten Engelen, Rogier de Haan, and Steven Woudenberg, all of whom have been supportive in their own ways. My paranymphs Dawa Ometto and Ruth Waumans deserve special mention for helping me pull through when the going was difficult.

The conversations with them, on and off topic, and their willingness to be my sounding board, have made a world of difference.

The steady, loving support from my family has been invaluable. I am deeply grateful to Jan Marten Kamphorst, Irma Jansen, and my mother, the late Corrie Kamphorst–van den Heuvel, as well as to extended family members Malcolm Richardson, the late Mary Richardson, and the late Peter Job. Each has shaped me and my life for the better. My gratitude also extends to my in-laws for their hands-on support in recent years and their continued interest in my well-being. I am deeply thankful for my son, Owen, for bringing so much joy into my world. Finally, I feel the deepest gratitude and appreciation for Arlette van Wissen, who has been by my side throughout the whole dissertation-writing process, from inception to conclusion, for better and for worse. I am so incredibly lucky to have such a strong partner in work, life and love. Thank you for everything.

Bart Anthony Kamphorst
Culemborg, February 2020

Curriculum Vitae

Bart Anthony Kamphorst was born on May 14th 1985 in Amersfoort, the Netherlands. He holds degrees in Philosophy (BA), Law (LLB, *with honours*) and Artificial Intelligence (MSc, *cum laude*), all obtained from Utrecht University. In 2009, during his MSc program, he was a visiting researcher at the A.I. Research Group at Harvard University. In 2014, while pursuing his PhD, he was a visiting scholar at Stanford University's Philosophy Department as well as a visiting researcher at Carleton University's Department of Psychology. From May 2016 to May 2019 he was a researcher in the Computational Intelligence Group at the Department of Computer Science of the Vrije Universiteit Amsterdam.

His broad academic interests are reflected in his publication record, with publications on subjects as diverse as human-computer teamwork, human-computer trust, (bedtime) procrastination, behavior change, motivation enhancement, and ethics of technology. Underlying this breadth of research is the belief that looking beyond disciplinary boundaries fosters mutual respect and understanding between researchers, enriches cultural and scientific discourses, and fuels progress towards addressing big questions in science and society.

For leisure, Bart enjoys ballroom dancing, scuba diving, and playing tennis. He currently resides in Culemborg with his partner, Arlette van Wissen, and their son Owen.

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 Suhrawardi and His Commentators* (dissertation), 2014.
volume 83. P.L. Lambertz, *Divisive matters Aesthetic difference and authority production 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 indenteitsdrager 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. Rajaiah, *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 vroegmodern 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 121. M.C. Göbel, *Human Dignity as the Ground of Human Rights. A Study in Moral Philosophy and Legal Practice* (dissertation), 2019.

volume 122. T. van 't Hof, *Enigmatic Etchings. True Religion in Romeyn de Hooghe's Hieroglyphica* (dissertation), 2019.

volume 123. M. Derks, *Constructions of Homosexuality and Christian Religion in Contemporary Public Discourse in the Netherlands* (dissertation), 2019.

volume 124. H. Nieber, *Drinking the Written Qur'an. Healing with Kombe in Zanzibar Town* (dissertation), 2020.

volume 125. B.A. Kamphorst, *Autonomy-Respectful E-Coaching Systems: Fending Off Complacency* (dissertation), 2020.

Quaestiones Infinitae

$\theta\pi$