Let's Do It!

Collective Responsibility, Joint Action, and Participation

We Gaan Ervoor!

Collectieve Verantwoordelijkheid, Gezamenlijk Handelen en Participatie

(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. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 18 mei 2018 des middags te 4.15 uur

door

Henricus Wilhelmus Anna Duijf

geboren op 3 oktober 1989

te Maasbree

Promotoren:	Prof. dr. D. Cohnitz
	Prof. dr. J-J. Ch. Meyer
Co-promotoren:	dr. J. M. Broersen
	dr. A. M. Tamminga

Cover painting by employees of Würth Nederland B.V. and Diederik Grootjans.

This research was funded by the European Research Council (ERC), ERC-2013-CoG project "Responsible Intelligent Systems" (REINS), No. 616512.

ISBN: 978-94-6103-066-5

Acknowledgements

The process that has led to the present dissertation has occasionally been difficult but mostly been fun. I am grateful to the guidance, freedom, and support that I have had the privilege to enjoy. Let me spend a few words to express my gratitude.

First and foremost, I would like to thank my supervisor Jan Broersen, whose influence has been most prominent and is greatly appreciated. I am extremely grateful for being able to join his fascinating research project titled 'Responsible Intelligent Systems'. We had many stimulating discussions on an extremely wide variety of topics. The freedom and trust you gave were valuable for this dissertation, in particular, and my personal development as a researcher, in general. I would also like to thank Allard Tamminga, to whom I owe a lot of academic maturity. He is always willing to dig into the details of mathematical or philosophical arguments and has helped me develop a high level of meticulousness. Special thanks go to my promotor John-Jules Meyer for his enthusiasm and unconditional trust. Although we don't meet as often as at the start of my project, I look back at our lively exchanges with great respect and joy.

The research project facilitated an open and healthy research environment. Besides my supervisors, I particularly thank Sjur Dyrkolbotn, Alexandra Kuncová, Tsz Yuen Lau, Niels van Miltenburg, Jesse Mulder, and Aldo Ramírez-Abarca for their invaluable suggestions and comments, and our inspiring conversations.

At the start of my PhD project, I was working in the Intelligent Systems group at the Computer Science department in Utrecht. I would like to thank my friendly colleagues for many fierce discussions on research and political issues, mostly during lunch, and would like to thank my fellow PhD candidates for jointly finding our way in the complex and perplexing academic environment, and for the occasional, welcome, distractions in the form of boardgamenights, karaoke parties, and Chinese hotpots. I greatly look forward to future editions.

In my second year, I moved to the Theoretical Philosophy group in Utrecht, where I quickly felt at home. In particular, I would like to thank the participants in the Analytical Philosophy seminar, who have helped me develop an eloquent writing style and an understanding of the research landscape in philosophy. More generally, the Attic Dwellers (PhD candidates and postdocs working in the attic) have provided a delightful atmosphere for fruitful discussions and revitalising interruptions, mostly in the form of a game of foosball. My academic life became more interesting by joining the PhD council of the Humanities faculty, whose aim is to represent and promote the interests of PhD candidates. I would like to thank my companions in the PhD council for this additional dimension.

Tenslotte wil graag de mensen bedanken die me het meest dierbaar zijn. Ik bedank graag mijn vrienden in Utrecht, Nijmegen, Maasbree en op verscheidene andere plekken in binnen- en buitenland voor de uiterst nuttige en cruciale ontspanning en de vele afleidende activiteiten die het leven kleurrijk maken.

Daarnaast wil ik graag mijn ouders bedanken voor de onvoorwaardelijke steun en opvoeding die onmiskenbaar hebben bijgedragen aan mijn vorming tot zelfstandig onderzoeker.

Bovenal wil ik Maxime bedanken voor de onbeschrijflijk liefdevolle steun.

Contents

This page intentionally contains only this sentence.

Introduction

Many aspects of our personal lives are shaped by interactions, such as when you buy a cup of coffee en route to work, navigate your way through traffic, organize a reading group with colleagues, and finally meet friends for a beer and card games in a pub. Additionally, but quite differently, collective decisions pervade and influence our personal lives to a large degree. This is shown by the elections of presidents in democratic societies, and the recent referendum result that caused Brexit, for example, but it is also illustrated by simply deciding on what to have for dinner together with your spouse. We are hence inevitably involved in what are called interdependent decision contexts, that is, scenarios in which the eventual outcome, and our evaluation thereof, depends on the interaction of several individuals.

These interdependent decision problems, especially in the form of committee decision-making and voting, are pervasive in our society. The overarching theme of the present work is to enquire about the relation between, on the one hand, collective decisions and, on the other hand, individual decisions. To clarify this focus, consider a version of the *discursive dilemma*: suppose a committee of academics, consisting of Marie, Mel, and Mo, is deciding on whether to award tenure to Mr Borderline. Imagine the university's tenure policy requires excellence in research, service, and teaching. Suppose the committee is to decide on the tenure by first voting on each of these fields of competence, then aggregating its members' votes on each of these fields by majority, and finally deriving the collective decision on

the tenure in line with the university's rules. If the members vote in accordance with Figure 1, then it turns out that they collectively decide to award tenure even though they are unanimously opposed.[1] Suppose that Mr Borderline turns out to be a bad candidate for tenure. Does it make sense to say that the group is collectively responsible for awarding tenure? Or, are any of the committee members individually responsible for contributing to awarding tenure?

	Research r	Service s	Teaching t	Tenure? $r \,\&\, s \,\&\, t$
Marie	Yes	Yes	No	No
Mel	No	Yes	Yes	No
Mo	Yes	No	Yes	No
Group	Yes	Yes	Yes	→Yes / ↓No

Figure 1: *The discursive dilemma.*

Do responsibility voids exist? That is, are there cases in which the group is collectively blameworthy for some outcome without any of its members being individually blameworthy for it? It is important to study the relation between collective and individual blameworthiness and, in particular, to characterize the conditions that must be met if responsibility voids are to exist. These issues become increasingly alarming once we are reminded of the way in which collective decisions shape our personal lives. Imagine you were competing with Mr Borderline for tenure. Can we justify a potential lack of individual blameworthiness for the faulty award of tenure? Moreover, since voting procedures are a vital element of our democratic institutions, these responsibility voids could reveal important weaknesses in democratic regulations. What if nobody can be blamed for policy implications regarding, for instance, your personal well-being, health, or prosperity? Finally, imagine you were a member of the committee that awar-

[1] The study of these cases is the subject of a strand of literature on social choice and judgement aggregation.

ded the tenure to Mr Borderline. What if you are mistakenly blamed for your involvement in the faulty award of tenure? We need a theory that helps us not to blame the wrong persons in such interdependent decision contexts.

In committee decision-making, the prospect of responsibility voids yields the question of whether there is a voting procedure and a combination of votes such that none of the members is an appropriate target of moral criticism, regardless of their involvement in bringing about a certain outcome. In interdependent decision problems, this prompts the question of whether there is a scenario involving a number of interacting individual agents and a combination of individual actions such that none of the individuals can be held morally responsible for contributing to a given state of affairs. If such cases exist, it is then paramount to classify the conditions under which responsibility voids may arise.

I aim to provide a systematic study of the relation between collective and individual blameworthiness and, in particular, of the conditions that must be met if responsibility voids are to exist. To do so, I rely on two central ideas. The *first idea* is that there is a world of difference between whether the decision problem is faced with friends or with foes. Simply think of this difference in a game of football: you approach your teammates in a different way from your opponents (at least when playing soccer, though perhaps not after the game). Analogously, it is common to distinguish between cases of cooperation and those of conflict. It may be hard to find out whether the people you interact with are partners or adversaries. Nonetheless, once their roles are recognized this shapes our sphere of interaction.

To bring this distinction to the fore, consider a version of the classic public goods dilemma, *the participation dilemma*: imagine a community has decided to set up a neighbourhood watch project to reduce crime rates. Let us assume that the project is successful if and only if a certain number of individuals participate in the project. Accordingly, contributing to the project means incurring a small

cost, but a successful project provides a benefit to each member of the community. The benefit for an individual thus depends on the success of the project and on whether she incurs a cost by contributing. What will or should the members of the community do?

Without delivering a full-blown theory at this point, the intuitive idea for contrasting the individual and community-directed perspective is to say that an individual asks herself 'What should *I* do?' whereas a group member first asks herself 'What should *we* do?' before determining how she may best contribute. Roughly stated, from an individual perspective, each will contribute only if she thinks it likely that she is pivotal for the success of the neighbourhood watch project. Alternatively, from a group-member perspective, it is reasonable that each will contribute unless she thinks it likely that her contribution is irrelevant for a successful project. So a group member is more positively disposed to contribute than an individualistic agent.

The *second idea* is that it is both common and vital to distinguish between different modes of acting when assessing individual moral responsibility. In criminal law, different modes of acting, that is, mental states that accompany the act, correspond with different levels of culpability. Although legal and moral responsibility need not align, I think that these modes of acting affect our attributions of moral responsibility and that studying them helps provide a systematic study of moral responsibility. The North American legal system uses the following distinctions regarding modes of acting, in decreasing order of culpability (taken from Dubber, 2002, pp. 60–80):[2]

- *Purposefully* – the actor has the 'conscious object' of engaging in conduct and believes and hopes that the attendant circumstances exist.

- *Knowingly* – the actor is certain that his conduct will lead to the result.

[2]See also Weigend (2014).

- *Recklessly* – the actor is aware that the attendant circumstances exist, but nevertheless engages in the conduct that a 'law-abiding person' would have refrained from.

- *Negligently* – the actor is unaware of the attendant circumstances and the consequences of his conduct, but a 'reasonable person' would have been aware.

- *Strict liability* – the actor engaged in conduct and his mental state is irrelevant.

The first two categories concern the intentions of the agent and what the agent knows about what she is doing, respectively. For example, suppose that Amy, a jealous wife, discovers that her husband is having a sexual affair with Vee. Wishing only to drive Vee away from the neighbourhood, she goes to Vee's house one night, rings the doorbell to check no one is home, pours petrol on the front door, and sets it on fire. Vee dies in the resulting fire. Amy is shocked and horrified. It did not occur to her that Vee might be physically in danger and there was no conscious plan in her mind to injure Vee when she began the fire. If Amy was charged with the murder of Vee, we may say that Amy neither purposefully nor knowingly saw to it that Vee died. In contrast, we can say that she purposefully and knowingly set fire to Vee's house.

Alternatively, we could amend the story such that Amy knew that Vee was present yet Amy only purposefully set fire to the house and did not intend to injure Vee. This would then constitute a difference between knowingly and purposefully doing something. Moreover, knowingly can indicate an awareness that certain circumstances exist, such as being in possession of a vicious dog. The owner need not intend to do harm, yet would still be liable for knowingly injuring others if his dog were to hurt them.

Nonetheless, it is plausible to say that Amy, by knowingly burning Vee's house, acted recklessly or negligently with regard to respecting human lives. It is

reasonable to say that Amy engaged in conduct that a law-abiding person would refrain from engaging in. Therefore, although she would not have knowingly or purposefully saw to it that Vee died, she may be said to have done so by reckless or negligent conduct.

These distinctions help to highlight important aspects of assigning responsibility and liability for conduct. In my systematic analysis of moral responsibility, I will therefore focus on several important components: *action*, *intention*, and *knowledge*.

In sum, these two central ideas yield two relevant dimensions for my study: modes of acting and different perspectives. The possible combinations are summarized in Figure 2. For example, the causal mode of acting yields a contrast between causal individual actions, collective causal actions, and contributory individual actions. Similarly, the group-member perspective induces a distinction between causally contributory actions, knowingly contributing, and intentionally participating.

Perspective / *Modality*	**Individual**	**Collective**	**Group member**
Causal	Causal action	Collective causal action	Causally contributory action
Knowingly	Knowingly doing	Collectively knowingly doing	Knowingly contributing
Intentionally	Intentionally doing	Collectively intentionally doing	Intentionally participating

Figure 2: *Modes of acting and different perspectives.*

In thumbnail form, the aim of the present work is this:

> It analyses the relation between collective and individual blameworthiness, and in particular conditions for the existence of responsibility

> voids, by taking seriously the idea that individual and member responsibility may vary and the idea that modes of acting link to different levels of moral responsibility.[3]

This finalizes the introductory sketch of the main research questions, the relevance of these topics, and the two central ideas that motivate my approach. Below I will describe the interdisciplinary viewpoint that is essential to the present work, provide an outline of what is to come, and end with a short guide to reading this thesis.

Method and Interdisciplinarity

The interdisciplinary method used in this work is, I think, one of its novel features.[4] The central disciplines are *philosophy*, *economics*, and *artificial intelligence*. I enjoy working across disciplines, finding inspiration in unexpected places, being able to draw together seemingly disconnected debates, and engaging with a variety of academics. Of course, this entails that some parts of the thesis may be more relevant and interesting to one discipline, whereas other parts are more central to another discipline.

To avoid confusion, let me briefly elaborate on the particular subfields of these three academic disciplines that I engage with. It should be clear that the relevant subfields of *philosophy* include action theory, moral responsibility, and collective

[3] In general, related work unwittingly conflates individual and member responsibility, only focuses on one mode of acting, or overlooks the different levels of culpability.

[4] The distinction between multidisciplinary, interdisciplinary, and transdisciplinary research may be vague. The common expressions for these are additive, interactive, and holistic, respectively. To clarify the concept of interdisciplinarity, Nissani (1997, p. 203 – emphasis in original) writes: "To begin with, a *discipline* can be conveniently defined as any comparatively self-contained and isolated domain of human experience which possesses its own community of experts. *Interdisciplinarity* is best seen as bringing together distinctive components of two or more disciplines. In academic discourse, interdisciplinarity typically applies to four realms: knowledge, research, education, and theory. Interdisciplinary knowledge involves familiarity with components of two or more disciplines. Interdisciplinary research combines components of two or more disciplines in the search or creation of new knowledge, operations, or artistic expressions. Interdisciplinary education merges components of two or more disciplines in a single program of instruction. Interdisciplinary theory takes interdisciplinary knowledge, research, or education as its main objects of study."

intentionality. Moreover, since collective intentionality plays a vital role in the foundations of the social sciences, my work may also be of interest to philosophers of the social sciences. It is, however, not my aim to contribute to normative ethics or epistemology as traditionally construed. That is, for instance, although I rely on some intuitive understanding of moral responsibility, I do not defend or position my intuitions with regard to ethical theories.

The *economics* literature on game theory provides an extremely helpful toolbox for analysing interdependent decision contexts, which are, as should be clear by now, my primary focus (von Neumann and Morgenstern, 1944; Schelling, 1960). Game and decision theory typically fall under microeconomic theory, in the company of, for example, supply and demand, and welfare economics. The interdisciplinary character of my perspective, however, justifies some departures from standard rational choice theory. More specifically, the work on epistemic game theory is vital for studying knowledge and action, and the team-reasoning account of cooperation will prove fruitful for studying joint action.

Although there are multiple branches of study within *artificial intelligence*, for the present work the most relevant one is the agent-oriented paradigm. To clarify, consider the following description by Nicholas Jennings:

> An *agent* is an encapsulated computer system that is situated in some environment and that is capable of flexible, autonomous action in that environment in order to meet its design objectives. (Jennings, 2000, p. 280 – emphasis added)[5]

Agent-oriented programming revolves around software agents, rather than revolving around objects as it does in object-oriented programming. One of the key differences is that agents are autonomous, in the sense that they have control over their behaviour, rather than the automatic nature of objects. Different types of

[5]See also (Wooldridge, 1997). To justify the relevance of this agent-oriented paradigm, Jennings (2000, p. 278) writes: "Agent-oriented approaches can significantly enhance our ability to model, design and build complex, distributed software systems."

agents are relevant for AI: intelligent agents (demonstrating some aspect of artificial intelligence), autonomous agents (capable of modifying the way of achieving an objective), and multi-agent systems (multiple agents that work together to achieve an objective). As such, the two central ideas that I have discussed are relevant for intelligently interacting artificial agents. For example, potential responsibility voids may point to a defect in the design of a multi-agent system: the existence of cases in which the system as a whole can be held collectively responsible although none of its individual artificial agents can be held individually responsible.

Despite its interdisciplinarity, the present work is foremost *philosophical*, or foundational, in nature and focuses on the philosophy of cooperation, joint action, and responsibility voids. The methodology should thus be viewed as a quest for arguments and justifications. For example, can we justify or rebut the idea that a joint action is reducible to a set of individual actions? One should therefore not expect any experimental findings or elaborate case studies here. My aim is to clarify the conditions that must be met if responsibility voids are to exist by systematically studying the interplay between various concepts, such as action, knowledge, and intention. The two central ideas help provide a detailed basis for studying the relation between collective decisions and individual decisions.

To aid and guide my theorizing, the practice of formal modelling is vital, that is, I try to provide mathematical models that are intended to represent relevant features and their interplay. This practice relies on a back-and-forth interaction between such models and educated intuitions. Our intuitions, for example regarding the interplay between action and knowledge, are often led astray in complex situations and a mathematical model could help iron out some confusion or obscurity. Conversely, our educated intuitions may justify appropriate

restrictions and extensions of certain models. The game-theoretical aspect of my study is, for example, best viewed as going beyond standard game-theoretical modelling because it includes concepts such as obligations and intentions.

Some say that mathematical models should be avoided as they are mere idealizations and fail to account for all the details of a particular case. This is valid criticism, but it is unclear to me whether other types of analyses can avoid this suspicion. Doesn't *any* explanation rely only on a partial description of a particular case? Psychologists typically focus on mentalistic constructs, sociologists on the social environment, and political theorists on institutional background. Don't they all provide useful information in particular cases? It is certainly vital to understand the extent and limitation of the application of our theories, including mathematical models. I therefore take care to design and motivate my models with an eye for the relevant literature.

A benefit of using formal models is that they are potentially extremely useful for generalizations. Let me give a simple illustration of this. Imagine that one would like to divide a number of oranges among one's siblings in a fair way. How should one do this? And under which conditions can one succeed in doing so? Well, a fair division would intuitively consist in each sibling getting an equal share of the total number of oranges. For example, if one has four siblings and twenty-four oranges, then a fair division would yield each getting six oranges. One way to find this answer is to employ the formal theory of arithmetic. That is, one simply calculates the result of $24/4$. We can similarly observe that it is sometimes impossible to find a fair division (for instance when one has four siblings and twenty-five oranges). The formal theory of arithmetic would therefore provide a full theory of a fair division of oranges among siblings and the conditions under which such a fair division can be found. Now, the sense of generality gained from using the formal theory of arithmetic is that the same formal theory can be used to study the fair division of apples among colleagues. Apples certainly differ from

oranges in substantial ways, just like your colleagues are markedly distinct from your siblings, but the formal theory of arithmetic can be used both for dividing oranges among siblings and for dividing apples among colleagues.

This generality of formal modelling helps bridge the playing field between philosophy, economics, and artificial intelligence. The models used in these disciplines share many characteristics and thereby enable fruitful synergies. The interdisciplinarity highlights that these formal models can be used for multiple purposes: to help clarify and understand philosophical doctrines and arguments, to aid microeconomic theorizing, and to provide a critical tool for designing and verifying complex agent-based artificial intelligence.

Outline

In this outline of the thesis I will highlight the crucial concepts, questions, and aims of the chapters.

In **Chapter 1** I introduce the basic mathematical foundations for the rest of the thesis. The present work relies on two formal traditions. The first is the tradition of logics for 'seeing to it that' (Belnap et al., 2001; Horty, 2001). One of the key philosophical concepts at play is that of *agency*. How, one may ask, can we make sense of the idea that agents may freely decide upon their course of action? To do so, I rely on the work of philosophical logicians synthesized in logics for 'seeing to it that', which are based on the central idea that the world is indeterministic. That is, at a particular moment in time, there may be several futures open. So, somewhat metaphorically, the future forks into different directions giving rise to alternative possibilities. The central idea for modelling agency is that an agent sees to it that X if and only if she performs an action whereby she constrains the possible futures to those where X holds.

The second formal tradition is the theory of games, as studied in the economics literature (von Neumann and Morgenstern, 1944). The significance of the theory

of games is now hard to overestimate. Its influence extends to sociology, psychology, philosophy, economics, mathematics, logic, and biology. Game theory can be viewed as the extension of decision theory to the realm of interdependent decision contexts. To clarify, decision theory typically deals with decision problems under certainty, risk, or uncertainty, whereas game theory studies interdependent decision problems. The theory of games is particularly helpful to highlight certain aspects of a scenario: the available individual actions, how combinations of individual actions yield outcomes, and preferences over these outcomes. The representation of such scenarios by way of a matrix allows for a clear and straightforward analysis of the underlying incentives and potential considerations. We will see that the notion of *agency* also takes central stage in game-theoretical models, as it naturally relates to what game theorists call α-effectivity.

In **Chapter 2** I study *the relation between collective obligations and individual obligations*. One might think that fulfilling individual obligations is necessary and sufficient for fulfilling collective obligations. Unfortunately, it is both unnecessary and insufficient. It may be that each member of a group fulfils her individual obligation without the group fulfilling its collective obligation, and vice versa. In particular, in interdependent decision contexts, it is possible that the group fails to fulfil its collective obligation while every member fulfils her individual obligation.

Given this discrepancy, I study whether the link can be re-established if the members regulate their individual actions via a *group plan*. That is, could the void in obligations be filled if the members of a group are able to communicate and agree upon a group plan? I therefore study member obligations that derive from agreeing on a group plan. But what makes a group plan a *good plan*, that is, a group plan that guarantees successful coordination of the group members' individual actions? It turns out that there are two key ingredients: optimality

and interchangeability. Briefly put, adopting a good group plan guarantees that whenever the collective obligation is not fulfilled then at least one member fails to fulfil its member obligation derived from the plan.

Finally, does the adoption of a group plan change the decision context? Arguably, if we adopt a plan together to cook risotto this evening, then it will not be ideal to disregard our agreement. I offer a model for studying how such agreements update the decision context. What is the relation between the individual obligations in the updated decision context and the member obligations that derive from the group plan? It turns out that when a good group plan is adopted, the member obligations that derive from the group plan coincide with the individual obligations in the updated decision context. Hence, if a good plan is adopted, then fulfilling individual obligations in the updated context guarantees that the combination of individual actions yields a group action that fulfils the collective obligation. Conversely, in the given circumstances, if the group fails to fulfil its collective obligation then at least one of its members fails to fulfil her individual obligation in the updated decision context.

In **Chapter 3** I focus on the interplay between *knowledge* and *action*. Instead of studying the causal effectivity of the agents, here I study the subjective abilities of the agents, which incorporate the information available to the agent. For example, an agent could be causally responsible for bringing about φ, while she does not know that she could bring about φ. What information needs to be absent in order to justify a lack of such practical knowledge? Or, conversely, what information is needed in order to justify such practical knowledge? I develop the idea that an individual agent knows how to φ if and only if there is a witness ψ such that she knows that she can φ by ψ-ing and she knows that she can knowingly ψ.

Analogously, I develop the idea that a group of agents collectively knows how to φ if and only if there is a refinement ψ which specifies each member's part, where, furthermore, they commonly know that they can φ by each member

playing her corresponding part and they know that each member knows how to play her part. I show that my theory of collective know-how adequately addresses coordination problems, that is, scenarios in which it is impossible for a group to collectively know how to coordinate. Moreover, I argue that both individual and collective know-how can be characterized using a simple epistemic extension of logics for 'seeing to it that'.

Knowing-how is important for the level of culpability associated with knowingly doing. This conception of moral responsibility is captured by the concept of subjective oughts. I therefore end the chapter with a reflection on the relation between subjective individual oughts and subjective collective oughts: does collectively knowingly failing to fulfil a collective obligation entail that some members are morally blameworthy? Why (not)?

In **Chapter 4** I study purposeful collective acts and their individual correlates by using *intentions*. Philosophers typically propose analysing joint action as an interlocking web of individual intentions (Bratman, 2014; Tuomela, 2005). How should we analyse these underlying individual intentions? Suppose that a collective intention, in the sense of a collective goal, has been adopted. Which objective should the group members adopt? I study three plausible candidates: a group member could adopt the group objective as her own; a group member may adopt as her objective performing an individual action that is compatible with a best group action; or, finally, a group member could adopt as her objective the realization of a best group action. These notions correspond to what I call pro-group intentions, team-directed intentions, and participatory intentions, respectively. One may think that there is no difference between these individual intentions; I show, however, that this is not the case: for example, the adoption of a pro-group intention produces different individual actions from the adoption of a participatory intention.

To study cooperation and joint action, I introduce intentions to game-theoretical models. The resulting models naturally relate to ideas from artificial intelligence, where artificial agents are commonly modelled as having (at least) three components: beliefs, desires, and intentions. Logical theories for such agents are often referred to as BDI logics (Cohen and Levesque, 1990; Rao and Georgeff, 1991; Meyer et al., 1999).

Team-reasoning theorists have argued that an adequate theory of collective intentionality should study the reasoning method endorsed by a group's members, rather than analysing the group members' mental states (Gold and Sugden, 2007). In a similar vein, some philosophers suggest that the individual intentions underlying joint action differ from standard individual intentions: the former are we-mode intentions, the latter are I-mode intentions (Hakli et al., 2010). Given these controversies, it seems paramount to meticulously study the individual intentions that underlie collective intentionality. I show that team-directed intentions naturally relate to we-mode intentions and team reasoning: they produce the same individual actions. My study thus bridges these paradigms.

Given these three different notions of individual intentions, we may ask which is best at producing successful cooperation. That is, for example, are there cases in which the adoption of pro-group intentions guarantees successful cooperation while the adoption of team-directed intentions fails to do so? Ultimately, I show that participatory intentions provide the best account of cooperation. That is, whenever a group of individuals strives to obtain some common objective, the best they can do is to each adopt a participatory intention.

Collective intentions and participatory intentions are important for the level of culpability associated with purposeful acts. There are at least three ways for a group to be collectively blameworthy: for not having performed a group action that fulfils its collective obligation; for not collectively intending to fulfil its collect-

ive obligation; or for collectively intending to refrain from fulfilling its collective obligation. In each of these cases, my theory of participatory intentions provides some insights into the grounds for holding a member individually responsible.

In **Chapter 5** I study potential responsibility voids by distinguishing between cooperation and competition. To do so, I rely on the reasoning-based analysis of cooperation offered by the team-reasoning account, which states that a set of individual actions is cooperative only if the individuals team reason rather than reasoning individualistically (Bacharach, 2006; Gold and Sugden, 2007). Accordingly, to apply this analysis to responsibility voids, I develop a sketch of a reasoning-based framework for responsibility. Despite the fact that reasoning falls beyond the levels of culpability, it will become clear that my reasoning-based framework takes into account all three modes of acting and categorizes the three distinct perspectives by way of reasoning schemas.

Does the existence of responsibility voids depend on whether the context is competitive or cooperative? If so, under which conditions can responsibility voids obtain? Although cooperative decision contexts may host responsibility voids, the conditions for the existence of such voids depend on the type of uncertainty the group faces: either external or coordination uncertainty.

In the conclusion, I summarize how the work done in the chapters provides a systematic study of the relation between collective and individual blameworthiness and, in particular, the conditions that must be met if responsibility voids are to exist. I close with some suggestions for future work that includes degrees of responsibility, hierarchical groups, problematic disagreement, and institutional responsibility.

A Reader's Guide

Chapter 1 establishes and introduces the central frameworks, viz. the theory of 'seeing to it that' and the theory of games, that are used and extended in the rest

of this thesis. Readers familiar with both these formal traditions may consider skipping that chapter; others are certainly encouraged to read it before venturing to the other chapters. In particular, § 1.3 may be of interest because it treats the conceptual and formal connections between these formal traditions.

The remaining chapters could each be read independently. There are of course links between the chapters, but they are mostly additional, not substantial, to a chapter's aims.

In Chapters 2–5, the *introductions* have three aims. First, to sketch the motivating questions and aims of the chapter. Second, to develop an informal view of the adopted approach. Third, to provide an outline of the chapter and its results. As such, the introductions are meant to help guide the reader to the parts of interest (and to perhaps help avoid the parts one dislikes).

The *conclusions* of Chapters 2–5 contain a discussion of the implications for the overall theme of this thesis: the existence of responsibility voids, and the relation between individual and collective moral responsibility.

Finally, the *proofs* of propositions, lemmas, observations, results, and theorems have been relegated to each chapter's appendix. Appendix E of Chapter 5 is an exception, as will be clear after reading the introduction of the corresponding chapter.

This page intentionally contains only this sentence.

1

Games and Agency

> *The most helpful invention of game theory for the social sciences is the payoff matrix.*
>
> Thomas Schelling (2010, p. 29)

> *We followed and extended the idea of treating agency as a modality – a modality that represents through an intensional operator the agency, or action, of some individual in bringing about a particular state of affairs.*
>
> Belnap, Perloff, and Xu (2001, p. 28)

To address the various problems in the scope of this dissertation, I rely on two formal traditions. The first tradition is the theory of games. The second tradition is a strand of literature on 'seeing to it that', abbreviated to STIT, which is concerned with action theory. In the final part of this introductory section I briefly discuss some formal and conceptual links between these two traditions.

[†]Most of what I cover in this introductory chapter is standard, although perhaps treated from an open-minded interdisciplinary perspective. I owe a great deal to Jan Broersen for introducing me to the theory of 'seeing to it that'. I also would like to thank Frederik Van De Putte and Allard Tamminga for our collaboration on the connection between games and STIT models (see Van De Putte, Tamminga, and Duijf, 2017).

1.1 An Introduction to the Theory of Games

Starting with the seminal work by von Neumann and Morgenstern (1944), the theory of games has been further developed and applied to study a wide range of phenomena and subjects. The theory provides a useful framework for thinking about interdependent decision problems (Schelling, 1960). That is, problems in which the outcome depends on the actions of several agents. To illustrate how the application of the framework might work in practice, I will start by discussing some well-known examples. Then I provide enough detail to set out a cogent introduction to the theory of games. It is therefore not my aim to provide an elaborate overview of all the theoretical and applied work done in the field of game theory. Rather, I will use and develop a framework to the degree necessary for my philosophical theorizing in later chapters.

Driving game. Two drivers approach one another in a two-lane road. Each can either keep left or keep right. They are both better off if they both choose the same respective side, otherwise they will have to stop, wasting precious time, or crash. We can use a simple matrix to describe this scenario; see Figure 1.1. (The number in the lower left corner represent Driver 1's utility and the number in the upper right corner represent Driver 2's utility.) Most importantly, this is a scenario in which the drivers must coordinate to solve the problem, there are many ways to solve the problem, and the drivers are indifferent about which way they do it.[1]

Many-hands problem. Two tourists are enjoying the sunny weather on the beach when each of them independently spots a drowning child in the ocean. Each can either go out to rescue the drowning child or decide to wait for the other to do the rescuing. They are both better off if at least one of them chooses to rescue the child,

[1] I take this description from Guala (2016, p. 24), who uses game-theoretical models to study institutions. Guala is inspired by Lewis (1969), who uses game-theoretical models for his philosophical analysis of conventions.

		Driver 2: Left	Driver 2: Right
Driver 1	Left	1 1	0 0
	Right	0 0	1 1

Figure 1.1: *Driving game.*

otherwise they will have wasted precious time, risking the success of a possible rescue operation; see Figure 1.2. Again, there are multiple ways of solving this coordination problem, and each is indifferent about the way they do it.[2]

		Tourist 2: Rescue	Tourist 2: Wait
Tourist 1	Rescue	1 1	1 1
	Wait	1 1	0 0

Figure 1.2: *Many-hands problem.*

Hawk-dove game. Think of two individuals in a state of nature who come into conflict over some valuable resource. To play dove is to offer to share the resource but to back down if the other attempts to take it all; to play hawk is to demand

[2]This game is meant to highlight some game-theoretical aspects of cases of overdetermination and the so-called bystander effect. The murder of Kitty Genovese has prompted research into the bystander effect (Gansberg, 1964; Darley and Latané, 1968). The game-theoretical structure of the case of the bystander effect is similar, although I will not investigate its psychological precursors. Braham and van Hees (2009) argue, on the basis of such problems of overdetermination, for a quantitative notion of causation. Later, they connect this work on causality to the distribution of responsibility in collective action problems (Braham and van Hees, 2012).

the whole resource, backed by a readiness to fight for it. We assume that fighting is costly for both parties and that the utility value of a half share of the resource is greater than half of the utility value of the whole; see Figure 1.3.[3]

		Player 2: Dove	Player 2: Hawk
Player 1	Dove	2, 2	3, 0
	Hawk	0, 3	-5, -5

Figure 1.3: *Hawk-dove game.*

These examples highlight the two fundamental components of a game-theoretical model: the game form and the utilities. A *game form* involves a finite set $N = \{i_1, \ldots, i_n\}$ of individual agents. Each individual agent i in N has a non-empty and finite set A_i of available individual actions.[4] I use a_i and a_i' as variables for individual actions in the set A_i. The Cartesian product $\times_{i \in N} A_i$ of all the individual agents' sets of actions gives the full set A of action profiles. I use a and a' as variables for action profiles in the set A.[5]

Definition 1.1 (Game Form)**.** *A* game form *S is a tuple $\langle N, (A_i) \rangle$, where N is a finite set of individual agents and for each agent i in N it holds that A_i is a non-empty and finite set of actions available to agent i. The set of action profiles A is given by $\times_{i \in N} A_i$.*

[3] I take this description from Gold and Sugden (2007, p. 111), who use game-theoretical insights to challenge prominent accounts of collective intentionality.

[4] In this thesis I will only use such finite game forms for my theorizing. There is, of course, a branch of game theory that deals with infinite games. The most famous example is the infinitely repeated prisoner's dilemma (see Axelrod, 1984).

[5] I adopt the notational conventions of Osborne and Rubinstein (1994, § 1.7) and omit braces if the omission does not give rise to ambiguities.

For each group $\mathcal{G} \subseteq N$ the set $A_{\mathcal{G}}$ of group actions that are available to group $\mathcal{G}$ is defined as the Cartesian product $\times_{i \in \mathcal{G}} A_i$ of all the individual group members' sets of actions. I use $a_{\mathcal{G}}$ and $a'_{\mathcal{G}}$ as variables for group actions in the set $A_{\mathcal{G}}$ $(= \times_{i \in \mathcal{G}} A_i)$. Moreover, if $a_{\mathcal{G}}$ is a group action of group $\mathcal{G}$ and if $\mathcal{F} \subseteq \mathcal{G}$, then $a_{\mathcal{F}}$ denotes the subgroup action that is $\mathcal{F}$'s component subgroup action of the group action $a_{\mathcal{G}}$. I let $-\mathcal{G}$ denote the relative complement $N - \mathcal{G}$. Finally, if $\mathcal{F} \cap \mathcal{G} = \emptyset$, then any two group actions $a_{\mathcal{F}}$ and $a_{\mathcal{G}}$ can be combined into a group action $(a_{\mathcal{F}}, a_{\mathcal{G}}) \in A_{\mathcal{F} \cup \mathcal{G}}$.

A *utility function* u assigns to each action profile a a value $u(a)$, and can be used to represent many different things. It is typically used by game and decision theorists to represent the preferences of an agent, or to represent the revealed preferences of an agent (Okasha (2016) provides a useful discussion on decision-theoretical interpretations of utility). But this is not the only available interpretation. Deontic logicians, for instance, use a binary utility function to represent a single moral code (see Hilpinen (1971), or, more specifically, Føllesdal and Hilpinen (1971, pp. 15–19)).[6] Depending on the interpretation of the utility function, derived game-theoretical notions should be interpreted differently. The value that an agent i's utility function u_i assigns to an action profile is usually given by a real number, which straightforwardly induces a comparison between action profiles, viz. a is more valuable than b according to u_i if and only if $u_i(a) > u_i(b)$. Depending on the interpretation of the utility function this means that (i) agent i prefers a over b, (ii) agent i always chooses a over b, or (iii) a is deontically better than b.[7]

Definition 1.2 (Game)**.** *A* game *S is a triple $\langle N, (A_i), (u_i) \rangle$, where $\langle N, (A_i) \rangle$ is a game form, and for each agent i in N it holds that u_i is a utility function that assigns to each action profile a in A $(= \times_{i \in N} A_i)$ a value $u_i(a) \in \mathbb{R}$.*

[6] Alternatively, in Chapter 4 I will use a binary utility function to represent the intention of an agent.

[7] One may ask whether it is always possible to represent an agent's preferences by a real-valued utility function. The most influential result is that of Savage (1954), who gives conditions under which it is possible to model an agent as if she were maximizing her expected utility, using a credence function and a real-valued utility function.

It may be helpful to point out that these games are generally called *normal-form* games (also sometimes called strategic-form games). These normal-form games can be taken to represent a situation in which several agents act simultaneously. These are contrasted with extensive-form games, which drop this simultaneity assumption and can be taken to represent sequential moves. It is important to note that each extensive-form game can be transformed into a normal-form game, although this transformation will remove the temporal structure.[8] Nonetheless, in this dissertation I will focus on normal-form games.

Game and decision theory deal with the following questions: 'What do people choose in certain decision problems?' and 'What should people choose in certain decision problems?'. The first question is descriptive and seeks to answer how people actually make decisions; the second is normative and studies how people should make decisions. In both enquiries game theorists have put forward many solution concepts. I will briefly discuss two of these that will be relevant throughout the dissertation.[9]

The *Nash equilibrium*, named after John Nash (1950, 1951), is perhaps the most well-known solution concept. Stated simply, Ann and Bob are in a Nash equilibrium if Ann is making the best decision she can, given Bob's actual decision, and Bob is making the best decision he can, given Ann's actual decision. Likewise, a group of agents are in a Nash equilibrium if each agent is making the best decision she can, given the actual decisions of the others. A Nash equilibrium is typically taken to represent a state in which no one has an incentive to deviate, given the choices of the others.[10]

[8] It may be interesting to note that normal-form games are typically assumed to imply complete, imperfect information, that is, the 'rules' of the game and the utility functions are commonly known and the agents cannot observe each other's simultaneous choice. In Chapter 3, however, I relax both these requirements by adding an epistemic dimension to normal-form games.

[9] The maximization of expected utility will not be discussed because my study will largely do without probabilities. Although this means that my investigations may not be fully general, for my purposes this gap will be irrelevant.

[10] Although the Nash equilibrium concept is widely accepted, it cannot be straightforwardly justified (Risse, 2000). Justifications based on epistemic conditions gave rise to the field of epistemic game theory (see Perea, 2012), originating from the work on rationalizability (Bernheim, 1984; Pearce, 1984).

Definition 1.3 (Nash Equilibrium)**.** *Let $S = \langle N, (A_i), (u_i)\rangle$ be a game. Then an action profile a is a Nash equilibrium if and only if for each agent i in N and for every $b_i \in A_i$ it holds that $u_i(a) \geq u_i(b_i, a_{-i})$.*

To illustrate, the Nash equilibria in the three discussed games are as follows: (left, left) and (right, right) in the driving game; (rescue, wait), (rescue, rescue), and (wait, rescue) in the many-hands problem; and (hawk, dove) and (dove, hawk) in the hawk-dove game.

Another intuitive principle is that of dominance, which comes in two guises: strict dominance and weak dominance. An action a_i *strictly dominates* action b_i in S, notation: $a_i \ll_S b_i$, if and only if a_i always yields a strictly better outcome than b_i, regardless of what the others do. An action a_i *weakly dominates* action b_i in S, notation: $a_i \preceq_S b_i$, if and only if a_i promotes the utility at least as well as b_i, regardless of what the other agents do. Weak dominance relates to Leonard Savage's "sure-thing principle"; he writes:

> I know of no other extralogical principle governing decisions that finds such ready acceptance. (Savage, 1972, p. 21)[11]

Definition 1.4 (Dominance)**.** *Let $S = \langle N, (A_i), (u_i)\rangle$ be a game. Let $a_i, a_i' \in A_i$ be individual actions available to i. Then*

$a_i \gg_S a_i'$ *iff for all* $a_{-i}'' \in A_{-i}$ *it holds that* $u_i(a_i, a_{-i}'') > u_i(a_i', a_{-i}'')$.

$a_i \succeq_S a_i'$ *iff for all* $a_{-i}'' \in A_{-i}$ *it holds that* $u_i(a_i, a_{-i}'') \geq u_i(a_i', a_{-i}'')$.

Strong dominance is defined in terms of weak dominance: $a_i \succ_S a_i'$ if and only if $a_i \succeq_S a_i'$ and $a_i' \not\succeq_S a_i$.

The set of *admissible individual actions* that are available to an individual agent i in a game S are defined in terms of the dominance ordering of A_i. An individual action a_i in A_i is admissible if and only if it is not strongly dominated by any individual action in A_i:

[11] In their axiomatic approach to decision theory, Luce and Raiffa (1957, see Section 13.3, and p. 306) express the admissibility requirement in Axiom 5 and write: "Axioms 1 through 5 seem quite innocuous and, so far as we are aware, all serious proposals for criteria satisfy them."

Definition 1.5 (Admissible Actions)**.** *Let $S = \langle N, (A_i), (u_i)\rangle$ be a game. Let i be an individual agent. Then the set of i's* admissible actions *in S, denoted by $Admissible_S(i)$, is given by*

$$Admissible_S(i) \quad = \quad \{a_i \in A_i : \text{ there is no } a_i' \in A_i \text{ such that } a_i' \succ_S a_i\}.$$

Admissibility captures the idea that an agent takes all actions of the other agents into consideration; none is entirely ruled out.[12] The admissibility concept has a long tradition in decision theory (see the discussion by Kohlberg and Mertens (1986, § 2.7)).

Note that this definition implies that there is at least one admissible action for each individual agent (a special case of Lemma B.1 in Appendix B). It may be useful to add that these three types of dominance are related in a straightforward way: strict dominance entails strong dominance, which entails weak dominance. To illustrate the admissibility requirement, let us reconsider the previously discussed games: in both the driving game and the hawk-dove game, any individual action is admissible, because neither weakly dominates the other; in the many-hands problem, only choosing rescue is admissible.

This concludes the brief introduction to the theory of games, showing its application in practice; highlighting its relevance to a range of topics; developing the basic framework; and discussing some solution concepts.

1.2 An Introduction to STIT Theory

The theory of 'seeing to it that', or simply STIT, has been developed in a series of papers by Belnap, Perloff, and Xu, culminating in their book (2001). It is a theory of agency that is cast against the background of branching time, which neatly

[12] Selten (1975) argues that even rational players, having made their choice, may with non-zero probability do something else by accident. In addition, Pearce (1984, Lemma 4) shows that an action is admissible if and only if it maximizes expected utility with respect to a probability function that assigns positive probability to every move of the opponent. This means that an expected utility maximizer should avoid inadmissible actions.

models the indeterministic nature of time. Although I will eventually settle for a different modelling, it will be useful to present the ideas of STIT theory within this branching-time semantics (§ 1.2.1) before giving my simplified version of it (§ 1.2.2).

1.2.1 Agency in Branching Time

The seminal contributions of Prior (1967) and Thomason (1970, 1984) gave rise to the theory of branching time that would later serve as the backbone for STIT semantics (Belnap et al., 2001; Horty, 2001). The branching-time models originate from a philosophical enquiry into the truth-values of temporal sentences, for example, so-called future contingencies. Belnap et al. (2001) present a detailed account of how our indeterministic world can be modelled.[13] The fundamental idea is to represent the world as moments ordered in a tree of histories. (It is important to note a possible confusion: 'histories' are taken to include future moments.) Figure 1.4 presents such a structure. The upward branching of histories represents the openness of the future. Although histories may branch at a particular moment, it is conceivable that there are moments at which no history branches. The absence of backward branching represents the determinateness of the past, that is, the fact that every moment has only a single past sequence of events. Each history in this tree-like structure represents a complete temporal evolution of the world.

A *branching-time model* involves a set of moments M, a set of histories $H \subseteq 2^M$, and a relation $<$ between moment/history pairs which represents the progression of events along a history. I use m, m' as variables for moments in M and h, h' as variables for histories in H. When a moment m and a history h satisfy $m \in h$, this can be taken to mean that m occurs on h, or that h passes through m. Because

[13] Perloff and Belnap (2011, pp. 583–584) write: "Part of the idea of indeterminism as we conceive it is that at any given moment there are a variety of ways in which the world might proceed. Such possibilities are real, not merely epistemic; they are possibilities."

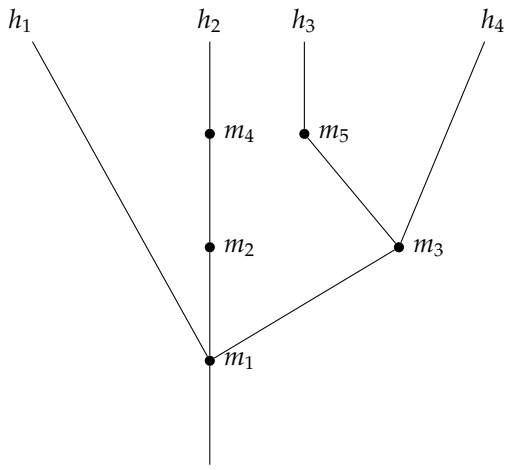


Figure 1.4: *A branching-time structure.*

of indeterminacy, there may be multiple histories that pass through m, so I let $H_m = \{h \in H \mid m \in h\}$ denote the set of histories through m. I use $\langle m, h\rangle$ as variables for moment/history pairs that satisfy $m \in h$ or, equivalently, $h \in H_m$. These moment/history pairs may be viewed as *dynamic states*, as opposed to static states; they include the current moment and the complete temporal evolution of the world. Finally, a valuation function V assigns to each propositional variable $p \in \mathcal{P}$ the set of dynamic states $V(p)$ where p obtains.

Definition 1.6 (Branching-time Model)**.** *A* branching-time frame *is a tuple* $BTF = \langle M, H, <\rangle$*, where* M *is the set of moments,* $H \subseteq 2^M$ *is the set of histories, and* $< \subseteq (M \times H)^2$ *is a relation on moment/history pairs* $\langle m, h\rangle$*. Moreover, BTF is required to satisfy the following:*

- *for every history* h*, the ordering* $<_h$ *on* h *induced by* $<$*, viz.* $m <_h m'$ *iff* $\langle m, h\rangle < \langle m', h\rangle$*, constitutes a linear ordering.*

(Linear Histories)

- *for every moment m and all histories h, h' such that $m \in h \cap h'$, it holds that $\{m' \in h \mid m' <_h m\} = \{m' \in h' \mid m' <_{h'} m\}$. In other words, there is a single past sequence of events.*

(Past Determinacy)

A branching-time model *$BTM = \langle M, H, <, V\rangle$ is a branching-time frame supplemented with a valuation $V : \mathcal{P} \to 2^{M \times H}$.*

These branching-time models are typically used to provide semantics for a logical language that includes a past operator P, a future operator F, and a historical necessity operator $\Box$.[14] Intuitively, $\mathsf{P}\varphi$ is true at a dynamic state $\langle m, h\rangle$ if and only if there is a moment m' before m on h where φ obtains; $\mathsf{F}\varphi$ holds at a dynamic state $\langle m, h\rangle$ if and only if there is a moment m' after m on h where φ obtains. The historical necessity operator $\Box\varphi$ expresses that φ holds at the current moment in the dynamic state $\langle m, h\rangle$ regardless of how the future unfolds. The dual $\Diamond\varphi$ expresses that there is a possible way for the future to unfold such that φ holds now.

Definition 1.7 (Evaluation Rules Temporal Formulas)**.** *Let $BTM = \langle M, H, <, V\rangle$ be a branching-time model and let φ be a formula constructed using propositional constants and temporal operators P, F, and $\Box$. Then the truth of φ, in a dynamic state $\langle m, h\rangle$ in BTM, notation: $BTM, \langle m, h\rangle \vDash \varphi$, is given by the following (suppressing the standard propositional clauses and the model BTM):*

$\langle m, h\rangle \vDash \mathsf{P}\varphi$ *iff* *there is an $m' \in h$ satisfying $m' <_h m$ and $\langle m', h\rangle \vDash \varphi$;*
$\langle m, h\rangle \vDash \mathsf{F}\varphi$ *iff* *there is an $m' \in h$ satisfying $m <_h m'$ and $\langle m', h\rangle \vDash \varphi$;*
$\langle m, h\rangle \vDash \Box\varphi$ *iff* *every $h' \in H_m$ satisfies $\langle m, h'\rangle \vDash \varphi$.*

[14] The literature also discusses other temporal operators; the most common additional operators are the next operator and the until operator.

Given these semantics, the idea that the future may still be open is represented by the invalidity of the formula $\mathsf{F}\varphi \rightarrow \Box\mathsf{F}\varphi$. In other words, there is a branching-time model BTM and a moment/history pair $\langle m, h\rangle$ such that $BTM, \langle m, h\rangle \vDash \mathsf{F}\varphi$ while $BTM, \langle m, h\rangle \nvDash \Box\mathsf{F}\varphi$. Or, equivalently, $BTM \vDash \mathsf{F}\varphi \wedge \Diamond\neg\mathsf{F}\varphi$.

The fundamental idea of agency in STIT theories builds on these branching-time models. At a particular moment m we may view H_m as representing the possibilities that are still open. Conversely, the histories outside H_m are no longer possible, or accessible, at moment m. Given that the histories in H_m are still open, an action, or choice, of an agent is viewed as restricting the possible histories to a subset K of H_m. Accordingly, 'the agent sees to it that φ' means that the truth of φ is guaranteed by an action or choice K of the agent. When Ann empties her glass of milk, the nature of her action on this view is to constrain the possible histories to those where the glass of milk is emptied. Hence, an action is identified with a subset of the possible histories. This induces the reading that an agent sees to it that φ only if she performs an action, thereby constraining the possible worlds to only φ-worlds.

A *branching-time agency model* supplements a branching-time model with a finite set of agents Ags and sets of available actions, one for each group of agents at each moment. Given a moment m and a group of agents $\mathcal{H}$, the set of available actions is given by a collection of subsets of the possible histories $Act^m_{\mathcal{H}} \subset 2^{H_m}$.[15] Since a history can be viewed as the complete temporal evolution of the world, it includes the actions that the agents are performing. The particular action that the group $\mathcal{H}$ executes at a dynamic state $\langle m, h\rangle$ is given by $Act^m_{\mathcal{H}}(h)$, which is the action $K \in Act^m_{\mathcal{H}}$ satisfying $h \in K$.[16]

[15] STIT theorists are ambiguous about the interpretation of *Act* as either choices or actions. According to Horty (1996, p. 274 – emphasis added), "*Act* is a device for representing the constraints that an agent is able to exercise upon the course of history at a given moment, the *actions* or *choices* open to him at that moment". Belnap et al. (2001, pp. 33–34 – notation adapted and emphasis added) writes: "The equivalence classes belonging to *Act* can be thought of as the possible *choices* or *actions* available." For my current purposes, the elements of *Act* are best thought of as actions.

[16] The fact that there is a unique action $K \in Act^m_{\mathcal{H}}$ satisfying $h \in K$ follows from the requirement that $Act^m_{\mathcal{H}}$ be a partition of H_m (see below).

Definition 1.8 (Branching-time Agency Model)**.** *A* branching-time agency frame *is a tuple* $BTAF = \langle M, H, <, Ags, (Act^m_{\mathcal{H}})\rangle$*, involving a branching-time frame* $\langle M, H, <\rangle$*, a finite set of agents* Ags*, and for each moment* m *and each group of agents* $\mathcal{H} \subseteq Ags$ *it holds that* $Act^m_{\mathcal{H}} \subset 2^{H_m}$ *is a finite set of actions available to group* $\mathcal{H}$ *at moment* m*, satisfying the following:*

- *for every moment* m *and every group* $\mathcal{H}$*,* $Act^m_{\mathcal{H}}$ *constitutes a partition of* H_m*.*
  (Partition)

- *for every moment* m*, every group* $\mathcal{H}$*, and all histories* $h, h' \in H_m$*, if there is a moment* m' *such that* $m' \in h \cap h'$ *and* $m <_h m'$*, then* $h \in Act^m_{\mathcal{H}}(h')$*.*
  (No Choice between Undivided Histories)

- *for every moment* m *and all groups* $\mathcal{F}, \mathcal{G}$*, if* $\mathcal{F} \subseteq \mathcal{G}$ *then* $Act^m_{\mathcal{F}} \supseteq Act^m_{\mathcal{G}}$*.*[17]
  (Agent Monotonicity)

- *for every moment* m*, all histories* h, h'*, and all groups* $\mathcal{F}, \mathcal{G}$*, if* $\mathcal{F} \cap \mathcal{G} = \emptyset$ *then* $Act^m_{\mathcal{F}}(h) \cap Act^m_{\mathcal{G}}(h') \neq \emptyset$*.*[18]
  (Independence of Agency)

A branching-time agency model *is a branching-time agency frame supplemented with a valuation* $V : \mathcal{P} \to 2^{M \times H}$*.*[19]

Figure 1.5 depicts such a branching-time agency frame. For example, at m_1, agents i and j both have two available actions, where each action is identified with a subset of the possible histories, in particular, $Act^{m_1}_i = \{\{h_1, h_5, h_6\}, \{h_2, h_3, h_4\}\}$. These branching-time agency models are used to interpret a logical language that

[17] It is important to note that I do not require these models to satisfy the *intersection property*, which states that $Act^m_{\mathcal{G}} = \bigcap_{i \in \mathcal{G}} Act^m_i$ (see Definition 1.12). The main reason not to have this requirement is to enable and simplify the proof of completeness.

[18] See Horty (2001, § 2.4) for a useful discussion of the independence of agency requirement.

[19] Branching-time models are often conceived as trees. However, for some purposes a branching-time model is better thought of as a forest consisting of several independent trees. This may, for instance, be important when adding an epistemic indistinguishability relation. For example, to allow for the possibility that an agent does not know the exact past, a branching-time model needs to be interpreted as a forest, rather than a tree.

includes agency operators $[\mathcal{H}\ \mathsf{stit}]$, one for each group $\mathcal{H}$. Intuitively, $[\mathcal{H}\ \mathsf{stit}]\varphi$ is true at a dynamic state $\langle m, h\rangle$ if and only if the truth of φ is guaranteed by the action of the group $\mathcal{H}$. That is, the group $\mathcal{H}$ performs an action K thereby constraining the possible histories to only those where φ holds. It may be useful to add that $[\mathcal{H}\ \mathsf{stit}]\varphi$ may be interpreted, relative to a dynamic state $\langle m, h\rangle$, as 'group $\mathcal{H}$ guarantees that φ holds regardless of what the others do'.[20]

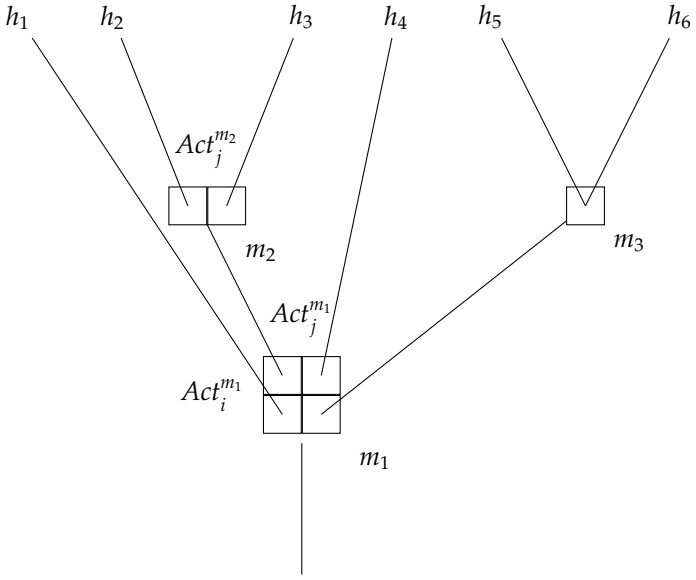


Figure 1.5: *A branching-time agency frame.*

Definition 1.9 (Evaluation Rule Agency)**.** *Let BTAM* $= \langle M, H, <, Ags, (Act^m_{\mathcal{H}}), V\rangle$ *be a branching-time agency model. Then the evaluation rule for the STIT operator is given by the following (suppressing the model BTAM):*

$\langle m, h\rangle \vDash [\mathcal{H}\ \mathsf{stit}]\varphi$ *iff for every* $h' \in Act^m_{\mathcal{H}}(h)$ *it holds that* $\langle m, h'\rangle \vDash \varphi$.

[20] The STIT operator is naturally connected to α-effectivity as used in game theory. I will discuss the relation between STIT theory and game theory in §1.3.

This concludes my discussion of the traditional STIT framework. To those familiar with STIT theory, note that the discussed operator is known as the Chellas STIT operator (named after Chellas, 1992). Other operators have been introduced and discussed in the literature, most importantly, the deliberative and the achievement STIT operator (originally proposed by Belnap and Perloff (1988), and later simplified by Horty and Belnap (1995)). STIT theory is related to logics of bringing it about (Elgesem, 1993, 1997); the difference is that STIT frameworks typically build on branching-time structures and impose the independence of agency requirement. STIT frameworks have been extended to include *strategic action* by Horty (2001, Chapter 7) and Broersen (2009) (see, for instance, Broersen and Herzig, 2015, §§ 3–6). The standard references for the beginnings of STIT theory are Belnap et al. (2001) and Horty (2001); for more recent surveys I would recommend Broersen and Herzig (2015) and Xu (2015).

1.2.2 STIT Models

In this section I will briefly introduce simplified models that are an abstraction of the branching-time agency models discussed in the previous section. These simplified models abstract away from the branching-time structure by using standard possible-world semantics. These models have been fruitfully applied to study meta-logical aspects of various STIT logics, such as completeness and axiomatization (see, for instance, Herzig and Schwarzentruber, 2008, §§ 5–6).

STIT models can be taken to represent possibilities and group actions at a single moment in time. These models abstract away from the temporal progression of the worlds. A STIT model involves a finite set of agents *Ags* and a set of possible *dynamic* worlds *W*, which can be viewed as the dynamic states based on a particular moment. I use the non-standard terminology of possible *dynamic* worlds to highlight the connection to dynamic states in branching-time models. It is important to note that a dynamic world is taken to include the complete

temporal evolution of the world.[21] Given a group of agents $\mathcal{H}$, the set of available group actions is given by a collection of subsets of the possible dynamic worlds $Act_{\mathcal{H}} \subset 2^W$. The particular action that the group $\mathcal{H}$ executes at a dynamic world w is given by $Act_{\mathcal{H}}(w)$, which is the action $K \in Act_{\mathcal{H}}$ satisfying $w \in K$.

Definition 1.10 (STIT Models)**.** *A* STIT frame *is a tuple $\langle W, Ags, (Act_{\mathcal{H}})\rangle$, involving a set of possible dynamic worlds W, a finite set of agents Ags and for each group of agents $\mathcal{H} \subseteq Ags$ it holds that $Act_{\mathcal{H}} \subset 2^W$ is a finite set of actions available to group $\mathcal{H}$, satisfying the following:*

- *for every group $\mathcal{H}$, $Act_{\mathcal{H}}$ constitutes a partitioning of W.*

  (Partitioning)

- *for all groups $\mathcal{F}, \mathcal{G}$, if $\mathcal{F} \subseteq \mathcal{G}$ then $Act^m_{\mathcal{F}} \supseteq Act^m_{\mathcal{G}}$.*

  (Agent Monotonicity)

- *for all dynamic worlds w, w' and all groups $\mathcal{F}, \mathcal{G}$, if $\mathcal{F} \cap \mathcal{G} = \emptyset$ then $Act_{\mathcal{F}}(w) \cap Act_{\mathcal{G}}(w') \neq \emptyset$.*

  (Independence of Agency)

A STIT model *is a STIT frame supplemented with a valuation $V : \mathcal{P} \to 2^W$. A* rich STIT model *is a STIT model supplemented with utility functions $u_i : W \to \mathbb{R}$, one for each agent $i \in Ags$.*[22]

The (rich) STIT models closely resemble *choice structures* and *consequentialist models* (Kooi and Tamminga, 2008, § 2), *STIT choice structures* (van Benthem and Pacuit, 2014, § 2), and *choice Kripke models* and *consequentialist choice Kripke models* (Ciuni and Horty, 2014, §§ 23.2–23.3).

[21] This aspect is *essential* and *non-standard*. The ramifications of this interpretation of dynamic worlds will be emphasised throughout the thesis. For example, it highlights that we can speak of the action performed by an agent at a dynamic world, whereas this would be fallacious or elusive for standard possible worlds.

[22] See the discussion of utility functions in game theory in § 1.1.

Figure 1.6 depicts a STIT frame. These STIT models can be used to interpret a logical language that includes the historical necessity operator $\Box$ and agency operators $[\mathcal{H}\ \mathsf{stit}]$, one for each group $\mathcal{H}$. The evaluation rules for these operators in STIT models mirror those for branching-time agency models (see Definitions 1.7 and 1.9). These models can therefore be used to provide semantics for the following logical language:

Definition 1.11 (Syntax)**.** *The formal language $\mathfrak{L}_{STIT}$ is as follows:*

$$\varphi \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\mathcal{H}\ \mathsf{stit}]\varphi,$$

where p ranges over a given countable set of propositions P, i ranges over a given finite set of agents Ags, and $\mathcal{H}$ ranges over subsets of Ags.

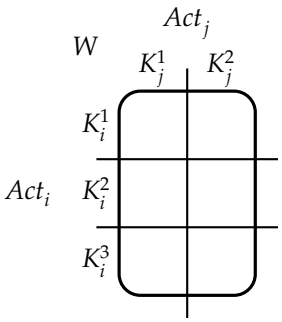


Figure 1.6: *A STIT frame.*

1.3 Connecting STIT Models and Games

The connection between STIT theory and the theory of games is simple: game-theoretical models correspond to a subclass of STIT models. Under this correspondence, an agent can see to it that some state of affairs is realized if and only if she is α-effective for that state of affairs. We need to do a bit of formal work in or-

der to fully grasp this correspondence. This connection has already been noticed and exploited by Horty (2001) to provide semantics for group obligations.[23] Others have subsequently further developed and exploited this analogy (see Turrini, 2012; Goranko et al., 2013; Tamminga, 2013; Ciuni and Horty, 2014; Bartha, 2014).

To establish a straightforward connection with games, two conditions for STIT models are crucial. First, the *intersection property* requires that any group action can be identified with a combination of individual actions. Second, *determinacy* requires that each group action of the grand coalition determines a unique dynamic world.

Definition 1.12 (Intersection Property & Determinacy)**.** *Let $\mathcal{F} = \langle W, Ags, (Act_{\mathcal{H}})\rangle$ be a STIT frame. We say that $\mathcal{F}$ satisfies the* intersection property *if and only if*

(IP) for every dynamic world w and for every group $\mathcal{H}$ it holds that $Act_{\mathcal{H}}(w) = \bigcap_{i\in\mathcal{H}} Act_i(w)$.

We say that $\mathcal{F}$ is deterministic *if and only if*

(Det) for every dynamic world w it holds that $Act_{Ags}(w) = \{w\}$.

We say that a STIT model or rich STIT model has the property – (IP) or (Det), respectively – if and only if the STIT frame it is based on has the corresponding property.

To state the correspondence theorem that connects STIT models to games, we need to elaborate on the notion of *effectivity for outcomes*. An outcome can be thought of as a (non-empty) set of possible dynamic worlds. Roughly speaking, an agent is effective for a given outcome in the model – STIT or game – if and only if she can ensure that outcome by some action, regardless of what the other agents do.

Definition 1.13 (Effectivity)**.** *Let $\mathcal{F} = \langle W, Ags, (Act_{\mathcal{H}}), (u_i)\rangle$ be a rich STIT frame and let $X \subseteq W$ represent an outcome. We submit that a group $\mathcal{H} \subseteq Ags$* is effective for *$X$ in $\mathcal{F}$ if and only if*

[23] My study in Chapter 2 relies on Horty's analysis of collective and individual obligations.

there is a $K \in Act_{\mathcal{H}}$ such that $K \subseteq X$.[24]

A similar definition can be given for games (recall Definition 1.2). Let $S = \langle Ags, (A_i), (u'_i)\rangle$ be a game and let $Y \subseteq A$ represent an outcome. We submit that a group $\mathcal{H} \subseteq Ags$ is effective for *Y in S if and only if*

there is an $a_{\mathcal{H}} \in A_{\mathcal{H}}$ such that $\{a' \in A \mid a'_{\mathcal{H}} = a_{\mathcal{H}}\} \subseteq Y$.[25]

Stated simply, we say that two models $\mathcal{M}_1$ and $\mathcal{M}_2$ – STIT or games – are *equivalent* if and only if for any given outcome X and every group of agents $\mathcal{H}$ it holds that $\mathcal{H}$ is effective for X in $\mathcal{M}_1$ if and only if $\mathcal{H}$ is effective for X in $\mathcal{M}_2$. The correspondence theorem states that every deterministic rich STIT frame that satisfies the intersection property corresponds to an equivalent game, and *vice versa*. (All claims are proved in Appendix A.)

Definition 1.14. *Let $\mathcal{F} = \langle W, Ags, (Act_{\mathcal{H}}), (u_i)\rangle$ be a deterministic rich STIT frame that satisfies the intersection property. The corresponding game $S^{\mathcal{F}} = \langle N, (A_i), (u'_i)\rangle$ is defined as follows:*

- $N := Ags$;
- *for every $i \in Ags$, $A_i := Act_i$;*
- $u'_i(a) := u_i(w)$, *where* $\{w\} = \bigcap_{i \in Ags} A_i$.

Because $\mathcal{F}$ is deterministic, a profile $a \in A$ can be identified with the world $w \in W$ that satisfies $\{w\} = \bigcap_{i \in Ags} A_i$. Likewise, a world $w \in W$ can be identified with $(Act_i(w))$. In the following we therefore loosely speak of W as profiles in $S^{\mathcal{F}}$.[26]

[24] Note that if X is represented by a proposition φ in a STIT model $\mathcal{M}$, then a group $\mathcal{H}$ is effective for X in $\mathcal{M}$ if and only if $\mathcal{M} \vDash \Diamond[\mathcal{H}\ \mathsf{stit}]\varphi$.

[25] It may be useful to add that this is equivalent to saying that there is an $a_{\mathcal{H}} \in A_{\mathcal{H}}$ such that $\{(a_{\mathcal{H}}, a'_{\_\mathcal{H}}) \mid a'_{\_\mathcal{H}} \in A_{\_\mathcal{H}}\} \subseteq Y$.

[26] It is important to note that this construction works because the rich STIT frame is assumed to satisfy (Det) and (IP). An indeterministic rich STIT frame would not allow for a straightforward identification of a world w with $\bigcap_{i \in Ags} A_i$. A violation of the intersection property undermines the condition in the corresponding game that $A_{\mathcal{H}} = \times_{i \in \mathcal{H}} A_i$.

Definition 1.15. *Let $S = \langle N, (A_i), (u'_i)\rangle$ be a game. The corresponding rich STIT frame $\mathcal{F}^S = \langle W, Ags, (Act_{\mathcal{H}}), (u_i)\rangle$ is defined as follows:*

- *$W := A$, $Ags := N$, and $u_i = u'_i$;*
- *for every $\mathcal{H} \subseteq Ags$ and for every $a \in A$, $Act_{\mathcal{H}}(a) := \{a' \in A \mid a'_{\mathcal{H}} = a_{\mathcal{H}}\}$.*

Note that $\mathcal{F}^S$ is deterministic and satisfies the intersection property.[27]

Theorem 1.1 (Correspondence Theorem)**.**

1. *Let $\mathcal{F} = \langle W, Ags, (Act_{\mathcal{H}}), (u_i)\rangle$ be a deterministic rich STIT frame that satisfies the intersection property. Then for the corresponding game $S^{\mathcal{F}}$ we have the following:*
   - *for every group $\mathcal{H}$ and every $X \subseteq W$ it holds that $\mathcal{H}$ is effective for X in $\mathcal{F}$ if and only if $\mathcal{H}$ is effective for X in $S^{\mathcal{F}}$.*
2. *Let $S = \langle N, (A_i), (u'_i)\rangle$ be a game. Then for the corresponding rich STIT frame $\mathcal{F}^S$, which is deterministic and satisfies the intersection property, we have the following:*
   - *for every group $\mathcal{H}$ and for every $Y \subseteq A$ it holds that $\mathcal{H}$ is effective for Y in S if and only if $\mathcal{H}$ is effective for Y in $\mathcal{F}^S$.*

Since I will later use logical machinery in my theorizing it may be helpful to restate this correspondence theorem in terms of a *logical* correspondence. On the one hand, this logical correspondence may help the reader familiar with game theory to understand what the STIT formalism is meant to capture. On the other hand, it helps, for instance, to connect my theories of individual and collective know-how to some discussions in game theory (in particular, see § 3.3.2). To establish this logical correspondence, we need to be able to interpret the logical

As an aside, it can be proven that an *indeterministic* rich STIT frame $\mathcal{M}_1$ that satisfies the intersection property can be transformed into a *deterministic* rich STIT frame $\mathcal{M}_2$ that satisfies the intersection property while preserving the effectivity for all groups $\mathcal{H} \subset Ags$, i.e. excluding the grand coalition (see my joint work Van De Putte, Tamminga, and Duijf, 2017).

[27] The proofs are straightforward: determinism follows from the fact that $A = \times_{i \in Ags} A_i$ and $Act_{Ags}(a) = \{a' \in A \mid a'_{Ags} = a_{Ags}\}$; the intersection property follows from the fact that $A_{\mathcal{H}} = \times_{i \in \mathcal{H}} A_i$.

language on games (Kooi and Tamminga, 2008; Tamminga, 2013). A game model is a game supplemented with a valuation. The evaluation rules for the modal operators are straightforward:

Definition 1.16 (Game Models)**.** *A* game model *is a tuple* $\langle N, (A_i), (u_i), V\rangle$ *involving a game* $\langle N, (A_i), (u_i)\rangle$ *and a valuation* $V : \mathcal{P} \to 2^A$*. The truth of a formula* $\varphi \in \mathfrak{L}_{STIT}$ *at a profile* a *in a game model* S*, notation:* $S, a \vDash \varphi$*, is given by the following (suppressing the standard propositional clauses):*

$S, a \vDash [\mathcal{H}\ \mathsf{stit}]\varphi$ *iff* *for every* $a' \in A$ *satisfying* $a_{\mathcal{H}} = a'_{\mathcal{H}}$ *it holds that* $S, a' \vDash \varphi$*;*

$S, a \vDash \Box\varphi$ *iff* *for every* $a' \in A$ *it holds that* $S, a' \vDash \varphi$*.*

Corollary 1 (Logical Correspondence)**.**

1. *Let* $\mathcal{M}$ *be a STIT model, and let* $w \in W$*. Then*
   - *for every formula* $\varphi \in \mathfrak{L}_{STIT}$ *it holds that* $\mathcal{M}, w \vDash \varphi$ *if and only if* $S^{\mathcal{M}}, w \vDash \varphi$*.*
2. *Let* S *be a game model, and let* $a \in A$*. Then*
   - *for every formula* $\varphi \in \mathfrak{L}_{STIT}$ *it holds that* $S, a \vDash \varphi$ *if and only if* $\mathcal{M}^S, a \vDash \varphi$*.*

This page intentionally contains only this sentence.

Appendix A

Games and Agency

Theorem 1.1 (Correspondence Theorem)**.**

1. *Let $\mathcal{F} = \langle W, Ags, (Act_{\mathcal{H}}), (u_i)\rangle$ be a deterministic rich STIT frame that satisfies the intersection property. Then for the corresponding game $S^{\mathcal{F}}$ we have the following:*

   - *for every group $\mathcal{H}$ and every $X \subseteq W$ it holds that $\mathcal{H}$ is effective for X in $\mathcal{F}$ if and only if $\mathcal{H}$ is effective for X in $S^{\mathcal{F}}$.*

2. *Let $S = \langle N, (A_i), (u'_i)\rangle$ be a game. Then for the corresponding rich STIT frame $\mathcal{F}^S$, which is deterministic and satisfies the intersection property, we have the following:*

   - *for every group $\mathcal{H}$ and for every $Y \subseteq A$ it holds that $\mathcal{H}$ is effective for Y in S if and only if $\mathcal{H}$ is effective for Y in $\mathcal{F}^S$.*

Proof. 1. Because $\mathcal{F}$ satisfies (Det) and (IP), each world w corresponds to the group action $Act_{Ags}(w)$, which is given by $\bigcap_{i\in Ags} Act_i(w)$, and therefore corresponds to the action profile $(Act_i(w))_{i\in Ags}$ in $S^{\mathcal{F}}$. The converse also holds: to each action profile a in $S^{\mathcal{F}}$ corresponds exactly one world, viz. the element of $\bigcap_{i\in Ags} a_i$. Since $K \in Act_i$ in $\mathcal{F}$ can be represented by a world w such that $K = Act_i(w)$, it holds that $K = \{w' \in W \mid Act_i(w') = Act_i(w)\}$, which corresponds to $\{a' \in A \mid a'_i = a_i\}$ in $S^{\mathcal{F}}$, where $K = a_i \in A_i$.

Let $X \subseteq W$, and let $\mathcal{H} \subseteq Ags$. Then the following are equivalent: (1) $\mathcal{H}$ is effective for X in $\mathcal{F}$, (2) there is a $K \in Act_{\mathcal{H}}$ such that $K \subseteq X$, (3) there are $K_i \in Act_i$, one for each $i \in \mathcal{H}$, such that $\bigcap_{i \in \mathcal{H}} K_i \subseteq X$, (4) there are $a_i \in A_i$, one for each $i \in \mathcal{H}$, such that $\{a' \in A \mid a'_i = a_i \text{ for each } i \in \mathcal{H}\} \subseteq X$, (5) there is an $a_{\mathcal{H}} \in A_{\mathcal{H}}$ such that $\{a' \in A \mid a'_{\mathcal{H}} = a_{\mathcal{H}}\} \subseteq X$, (6) $\mathcal{H}$ is effective for X in $S^{\mathcal{F}}$.

2. Follows immediately from unravelling the definitions and noting that a group action $a_{\mathcal{H}} \in A_{\mathcal{H}}$ in S corresponds to the group action $\{a' \in A \mid a'_{\mathcal{H}} = a_{\mathcal{H}}\} \in Act_{\mathcal{H}}$ in $\mathcal{F}^S$. □

Corollary 1 (Logical Correspondence)**.**

1. *Let $\mathcal{M}$ be a STIT model, and let $w \in W$. Then*
   - *for every formula $\varphi \in \mathfrak{L}_{STIT}$ it holds that $\mathcal{M}, w \vDash \varphi$ if and only if $S^{\mathcal{M}}, w \vDash \varphi$.*
2. *Let S be a game model, and let $a \in A$. Then*
   - *for every formula $\varphi \in \mathfrak{L}_{STIT}$ it holds that $S, a \vDash \varphi$ if and only if $\mathcal{M}^S, a \vDash \varphi$.*

Proof. Can be proven by standard induction on the complexity of the formula using the correspondences alluded to in the proof of Theorem 1.1. □

2

Collective Obligations, Group Plans, and Individual Actions

> *Of the major disciplines concerned with social behaviour, game theory alone, it seems, has eschewed, and got along to date without, group notions.*
>
> Michael Bacharach (2006, p. 72)

2.1 Introduction

Individual and collective obligations do not match: the fulfilment of a collective obligation is neither necessary nor sufficient for the fulfilment of individual obligations. It may be problematic when a collective failure to fulfil a collective obligation cannot be attributed to a member's failure to fulfil her individual obligation. This opens the possibility for voids between individual and collective responsibility. If group members aim to fulfil a collective obligation, they must act in such a way that the composition of their individual actions amounts to a group action that fulfils the collective obligation. We use a game-theoretical formalism

[†]This chapter is largely based on published work with Allard Tamminga (Tamminga and Duijf, 2017).

to study a strong sense of joint action in which the members of a group, using team reasoning, design and then publicly adopt a group plan. By highlighting particular group actions, a group plan specifies the individual actions that are the components of these highlighted group actions and thus specifies for every group member what she ought to do to contribute to the group's fulfilling its collective obligation. The public adoption of a group plan changes the context in which group members and other agents make a decision about what to do.[1] We give necessary and sufficient conditions under which a group plan that is designed to fulfil a collective obligation is a *good plan*, that is, it successfully coordinates the individual actions of the group members. Our central theorem concerns what happens if such a good plan is publicly adopted. We show that if the group members publicly adopt a good group plan, then the decision context is changed in such a way that for every group member it holds that she acts according to the group plan if and only if she performs an action that is one of the best things she can do in the changed decision context, regardless of the actions taken by all the others.

We illustrate our game-theoretical formalism with concepts and ideas from the philosophical and economics literature on deontic logic, team reasoning, collective intentionality, and joint action. This serves two purposes. On the one hand, the concepts and ideas from philosophy and economics help to clarify what our formalism is meant to model and thereby provide a partial conceptual justification for our formal analysis of the relation between collective obligations, group plans, and individual actions. On the other hand, we submit that our formalization captures important aspects of concepts and ideas that have been developed in the philosophical and economics literature and that therefore our formal results are of central relevance to the debate on team reasoning, collective intentionality, and

[1] Van Hees and Roy (2008) and Roy (2009a,b) present game-theoretical studies of how individual intentions might change the decision context for individual agents. Although their decision contexts include the intentions and actions of other agents, their approach differs from ours in that team reasoning, collective intentions, and group actions play no role in their analyses.

joint action. To illustrate this, we briefly discuss the relation between collective and individual rationality (§ 2.3) and the relation between collective and individual blameworthiness (§ 2.6) from the perspective of our formalism. We do not, however, engage in a prolonged philosophical discussion with the main protagonists in the debate, since our current purpose is to present our game-theoretical formalism and our results on the relation between collective obligations, group plans, and individual actions.

The chapter is set out as follows. In § 2.2, we define individual and collective obligations in terms of deontically admissible actions in a (single-shot) deontic game involving a single deontic ideality function. An individual agent is said to fulfil her *individual obligation* if and only if she performs one of her deontically admissible individual actions. Likewise, a group fulfils its *collective obligation* if and only if it performs one of its deontically admissible group actions. (For readability, we drop the adverb 'deontically' in 'deontically admissible'.) We show that the fulfilment of a collective obligation is neither necessary nor sufficient for the fulfilment of individual obligations. In § 2.3, we define a *group plan* as a set of group actions that are available to the group as a whole. Given such a group plan, a group member fulfils her *member obligation* specified by the plan if and only if she performs an action that is her component action of one of the group actions in the plan. We model the public adoption of a group plan as an update of the deontic ideality of the action profiles in a deontic game. In § 2.4, we give necessary and sufficient conditions under which a group plan is a *good plan*, that is, conditions under which a group plan guarantees that if every group member fulfils her member obligation specified by the plan, then the group itself fulfils its collective obligation. Our central theorem is proved in § 2.4.3: if a deontic game is updated with a good plan, then for every individual group member it holds that she fulfils her member obligation specified by the plan if and only if she fulfils her individual obligation in the deontic game that results

from updating the original deontic game with the plan. Consequently, if a group publicly adopts a good plan, then a collective failure to fulfil a collective obligation is always due to an individual failure to fulfil a member obligation specified by the plan, or equivalently, an individual failure to fulfil an individual obligation in the deontic game that results from updating the original deontic game with that plan. A brief discussion of the relation between collective and individual rationality follows. In § 2.5, we show that almost all our findings on deontic games transfer to what we call 'zero-preserving cooperation games' involving group-relative utility functions. On the basis of our discussion of cooperation games we then compare our approach to cooperation with the team-reasoning account of cooperation. In the concluding section, we use our formal analysis of collective obligations, group plans, and individual actions to study some logical aspects of backward-looking collective moral responsibility.

2.2 Individual and Collective Obligations

We use *deontic games* to study relations between collective obligations, group plans, and obligations of individual agents.[2] (In § 2.5, we investigate these relations using 'cooperation games'.) A deontic game is a particular type of game. (See § 1.1 for a more detailed introduction to the theory of games and further notational conventions.) It consists of a game form and a *deontic ideality* function d, which assigns to each action profile a in A a value $d(a)$ that is either 1 (if a is deontically ideal) or 0 (if a is not deontically ideal). We take deontically ideal action profiles to represent a single moral code, similar to deontically ideal worlds in the possible-worlds semantics for standard deontic logic (Hilpinen, 1971, pp. 13–15).[3] This

[2] Deontic games are similar to Schelling's (1960, p. 84) *pure-collaboration games* and Bacharach's (2006, p. 122) *coordination contexts*.

[3] Deontic logicians study the logical aspects of of normative expressions, like obligations, duties, permissions, right, and other related expressions. The seminal work by von Wright (1951) has sparked the field of deontic logic. Kanger (1971) and Anderson (1958) give semantical interpretations of deontic logic using deontically ideal worlds that represent what "morality prescribes" (Hilpinen, 1971, p. 21).

binary ordering of the action profiles in terms of deontic ideality can also be taken to reflect a simple preference ordering of agents who classify action profiles unanimously as 'good' or 'bad'.[4]

Definition 2.1 (Deontic Game)**.** *A* deontic game *S is a triple $\langle N, (A_i), d\rangle$, where $\langle N, (A_i)\rangle$ is a game form and d is a deontic ideality function that assigns to each action profile a in A ($= \times_{i\in N} A_i$) a value $d(a) \in \{0, 1\}$.*

To prove some of our claims about collective obligations, group plans, and obligations of individual agents, we must rule out deontic games in which no action profile is deontically ideal. Such deontic games are *flat*:

Definition 2.2 (Flat)**.** *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Then S is* flat *if and only if for all action profiles $a \in A$ it holds that $d(a) = 0$.*

We follow John Horty's (1996; 2001) deontic logic of agency and analyse a group's collective obligations in terms of its admissible group actions that are in turn defined using a dominance ordering of the group actions that are available to the group. Intuitively, group action $a_{\mathcal{G}}$ *weakly dominates* group action $a'_{\mathcal{G}}$ in deontic game S (notation: $a_{\mathcal{G}} \succeq_S a'_{\mathcal{G}}$) if and only if $a_{\mathcal{G}}$ promotes deontic ideality in S at least as well as $a'_{\mathcal{G}}$, regardless of what the group's non-members do.

Definition 2.3 (Dominance)**.** *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ be a group of agents. Let $a_{\mathcal{G}}, a'_{\mathcal{G}} \in A_{\mathcal{G}}$ be actions available to group $\mathcal{G}$. Then*

$$a_{\mathcal{G}} \succeq_S a'_{\mathcal{G}} \quad \textit{iff} \quad \textit{for all } a''_{\text{-}\mathcal{G}} \in A_{\text{-}\mathcal{G}} \textit{ it holds that } d(a_{\mathcal{G}}, a''_{\text{-}\mathcal{G}}) \geq d(a'_{\mathcal{G}}, a''_{\text{-}\mathcal{G}}).$$

Strong dominance is defined in terms of weak dominance: $a_{\mathcal{G}} \succ_S a'_{\mathcal{G}}$ if and only if $a_{\mathcal{G}} \succeq_S a'_{\mathcal{G}}$ and $a'_{\mathcal{G}} \not\succeq_S a_{\mathcal{G}}$.

[4] The following definitions are similar to the introductory definitions 1.1, 1.2, 1.4 and 1.5 yet they concern deontic games and they include *group* actions.

We define the set of *admissible group actions* that are available to a group $\mathcal{G}$ of agents in a deontic game S in terms of the dominance ordering of $A_{\mathcal{G}}$. A group action $a_{\mathcal{G}}$ in $A_{\mathcal{G}}$ is admissible if and only if it is not strongly dominated by any group action in $A_{\mathcal{G}}$:

Definition 2.4 (Admissible Actions)**.** *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ be a group of agents. Then the set of $\mathcal{G}$'s* admissible actions *in S, denoted by $Admissible_S(\mathcal{G})$, is given by*

$$Admissible_S(\mathcal{G}) \quad = \quad \{a_{\mathcal{G}} \in A_{\mathcal{G}} : \textit{ there is no } a'_{\mathcal{G}} \in A_{\mathcal{G}} \textit{ such that } a'_{\mathcal{G}} \succ_S a_{\mathcal{G}}\}.$$

Note that this definition implies that for every non-empty group of agents there is at least one admissible action (see Lemma B.1 in Appendix B). Since there always is at least one admissible action, in particular for any non-empty group and any of its non-empty subgroups, it makes sense to enquire about the relation between admissible group actions and admissible subgroup actions. First, note that an admissible group action is not necessarily composed of admissible subgroup actions. In the deontic game S_1 of Figure 2.1 it holds that (a'_i, a'_j) is admissible for the group $\{i, j\}$, although a'_i is not admissible for agent i and a'_j is not admissible for agent j. (Notice the resemblance with the many-hands problem of Figure 1.2.)

	a_j	a'_j	a''_j
a_i	1	1	1
a'_i	1	1	0
a''_i	1	0	0

Figure 2.1: *Deontic game S_1.*

Second, the composition of admissible subgroup actions is not necessarily an admissible group action. Deontic game S_2 of Figure 2.2 shows the deontic

analogue of the driving game (Figure 1.1). In S_2 it holds that a_i is admissible for agent i and a'_j is admissible for agent j, although (a_i, a'_j) is not admissible for the group $\{i, j\}$.

	a_j	a'_j
a_i	1	0
a'_i	0	1

Figure 2.2: *Deontic game* S_2.

Finally, we say that a group $\mathcal{G}$ fulfils its *collective obligation* in deontic game S if and only if it performs one of its admissible group actions. (Likewise, an agent i fulfils her *individual obligation* in deontic game S if and only if she performs one of her admissible individual actions.) Our discussion of Figures 2.1 and 2.2 shows that collective obligations and individual obligations do not match: the fulfilment of a collective obligation is neither necessary nor sufficient for the fulfilment of individual obligations. For example, this highlights that the driving game and the many-hands problem can give rise to cases where individual and collective obligations mismatch (Figures 1.1 and 1.2). To fulfil its collective obligation the group's members must ensure that their individual actions amount to an admissible group action. The gap between collective obligations and individual agency is filled by *member obligations*. Intuitively, a member obligation is what a group member ought to do in order to help ensure that the group fulfils its collective obligation, that is, in order to help ensure that the group performs an admissible group action. In this chapter we consider member obligations as following from a collective obligation via a group plan.

2.3 Group Plans and Member Obligations

A group that aims to fulfil a collective obligation has a coordination problem: the group members must cooperate to ensure that the composition of their individual

actions amounts to an admissible group action. In contrast to most coordination theorists, we assume that group members are able to communicate and to agree on a *group plan* to solve their coordination problem.[5] By agreeing on a group plan, it becomes common knowledge among the members of the group that the plan has been adopted. A group plan, once agreed upon, coordinates the individual actions of the members of the group: it allows them to act *simultaneously* and *unconditionally*, in the full belief that every group member acts according to plan. Group actions that are regulated by a group plan constitute a strong sense of joint action.[6] Raimo Tuomela calls this sense of acting together *proper joint action*:

> Acting together involves sociality in the relatively strong sense that such action must be based on joint intention or shared collective goal. This makes any case of acting together cooperative at least to the extent that the persons are collectively committed to making true a certain state of affairs. [...] In the strongest sense of acting together we require acting on a joint, agreed-upon plan. This I will call proper joint action. (Tuomela, 2000b, p. 7)

To design a group plan the group members must discover which are the best courses of action they can take *as a group*. To do so, group members must abandon their individual perspective and reason from the standpoint of a group and the group actions that are available to it. This kind of reasoning is known as *team reasoning* (Bacharach, 1999, 2006; Sugden, 1993, 2000, 2003) or *we-reasoning*

[5] Coordination theorists typically address the following problem: given a one-shot two-player game where no communication between the agents is possible, how are we to explain an agent's ability to anticipate the other agent's actions? Schelling (1960, pp. 54–58), Lewis (1969), Gauthier (1975), Sugden (1993, 1995, 2003), and Bacharach (2006) all study this type of coordination problem, arguing that concepts like *salience*, *focal points*, or *framing* are the key to understanding the coordination abilities of non-communicating agents.

[6] Weaker senses of joint action are explored by, for instance, Jackson (1987), Kutz (2000), and Chant (2007).

(Tuomela, 2013).[7,8] We take it that team reasoning permits the group members to come up with and agree on a group plan.[9] Once adopted, a group plan provides "a filter on options that are potential solutions" (Bratman, 1987, p. 35) to a group's coordination problem. By filtering out group actions, a group plan specifies the remaining group actions and in consequence it highlights the individual actions that are the components of these highlighted group actions. Accordingly, a group plan specifies for each group member what she ought to do.

We intend to understand how a group plan gives rise to member obligations by way of Margaret Gilbert's account of joint commitments. We submit that by adopting a group plan, a group of agents enters a joint commitment to execute the plan.[10] After entering such a joint commitment, the group members "owe each other conformity to the commitment. Thus, they have obligations towards each other" (Gilbert, 2006a, p. 156). Such obligations of joint commitment are genuine obligations, Gilbert argues: they meet the condition that "one who has an obligation to perform some action will have sufficient reason for performing it, sufficient reason that is independent of his own inclinations or self-interest and that cannot be eradicated by his own fiat" (Gilbert, 2006a, p. 157). Nonetheless, obligations of joint commitment are social rather than moral obligations.

The relation between a joint commitment and its concomitant social obligations is, Gilbert claims, *a priori*: the mere fact of entering a joint commitment implies social obligations on the members of the group.[11] They "owe each other actions by

[7] Bacharach (2006, p. 121) writes: "Roughly, somebody 'team-reasons' if she *works out the best feasible combination of actions for all the members of her team, then does her part in it*." Bacharach and Sugden assume that team reasoning leads to a unique group action. In the present setting we do not need this methodological assumption. In § 2.5.1 we compare our account of cooperation with Bacharach's and Sugden's team-reasoning account.

[8] Chapters 4 and 5 include more detailed discussions of team reasoning and we-reasoning.

[9] Similarly, Gold and Sugden (2007, p. 126) argue that "it is natural to regard the intentions that result from team reasoning as collective intentions".

[10] Compare Tuomela (2000b, p. 8): "In general, the performance of a joint action can be regarded as agreement-based if the plan has been accepted by the participants and if they have communicated their acceptances appropriately to the others so that a joint commitment to perform the joint action has come about."

[11] See also Tuomela (2005, p. 345): "If agreement making is in question, there will also be a publicly existing social (or, if you like, '*quasi*-moral') obligation to participate in joint action. This entailment of

virtue of their participation in the joint commitment *and that alone*. Consideration of what bad consequences might flow from violation of the commitment, for instance, is not relevant to the issue" (Gilbert, 1999, p. 151). A publicly adopted plan thus generates member obligations, regardless of whether or not the group members correctly assume that it is a good plan.[12] Robert Sugden concurs:

> To act as a member of the team is to act as a *component* of the team. It is to act on a concerted plan, doing one's allotted part in that plan without asking whether, taking other members' actions as given, one's own action is contributing towards the team's objective. [...] It must be sufficient for each member of the team that the plan itself is designed to achieve the team's objective: the objective will be achieved if everyone follows the plan. (Sugden, 1993, p. 86)[13]

We suggest that a group plan $P_{\mathcal{G}}$ of a group $\mathcal{G}$ in deontic game S be thought of as any subset of the set $A_{\mathcal{G}}$ of group actions that are available to $\mathcal{G}$ in S. By adopting a group plan, the group members enter a joint commitment to perform one of the group actions in the plan. Given an adopted group plan $P_{\mathcal{G}}$ and an individual group member i of $\mathcal{G}$, agent i fulfils her member obligation specified

an obligation can be regarded as a conceptual truth about the notion of agreement." Bacharach (2006, p. 64 – notation modified) is not so sure about the a priori character of this entailment relation: "[E]ven granted that an agent should decide that a certain profile a should be realized, why does this give her a reason to do her part in a?" Scanlon (1998, p. 317) argues against its a prioricity: "If a convention or social practice is taken to consist in the fact that people accept certain rules or norms and typically act in accordance with them, then we need a mediating moral principle to explain how such practices can be morally binding and generate specific obligations."

[12] See again Tuomela (2000b, p. 207): "If some persons make an agreement to cooperate, that entails the obligation to cooperate, be cooperation rational or not. Accepting such an obligating agreement in a full sense entails for the participants a collective commitment to fulfil it."

[13] Compare Sugden (2003, p. 72): "A cooperative morality enjoins each individual to *do her part* in achieving outcomes that are good for all. [...] [T]he individual does not ask whether her own actions, considered in isolation, yield preferred outcomes."

by the plan $P_{\mathcal{G}}$ if and only if she performs an action that is her component action of one of the group actions in the plan, that is, if and only if she performs an action from the set $\{a_i : a_{\mathcal{G}} \in P_{\mathcal{G}}\}$.[14]

2.3.1 Updating Deontic Games by Group Plans

We have argued that the adoption of a group plan generates member obligations and that these member obligations are specified by the group plan. There is, however, a second way in which the adoption of a group plan generates obligations. The adoption of a group plan can also be understood as inducing a change of the deontic ideality of the action profiles in a deontic game.[15] Because a change in deontic ideality might affect the dominance ordering of an agent's set of actions, such a change might lead to a different set of actions with which an agent, whether individual or group, might fulfil her individual or its collective obligation.

Technically, we model the adoption of a group plan in deontic game S as an *update* of the deontic ideality function of the game's action profiles. We use $S{\uparrow}P_{\mathcal{G}}$ to denote the deontic game that results from updating deontic game S with a group plan $P_{\mathcal{G}}$. To keep the update of the deontic ideality function technically manageable, we require that the group plan be adopted *publicly*. We may hence assume that after the public adoption of a group plan it is *common knowledge* among the agents in the game, whether or not they are members of the group that adopts the plan, that the plan has been adopted. The public adoption of a group plan hence changes the decision context not only for group members, but also for agents who are not in the group. After the public adoption of a group plan each agent in the outgroup decides what to do, on the supposition that the

[14] It may be useful to point out that this conception of member obligations naturally relates to important ideas in Chapter 3 (in particular, the discussion leading up to Theorem 3.4 and the characterization of collective know-how in Corollary 2) and Chapter 4 (in particular, the team-directed intentions discussed in § 4.4.2).

[15] Tuomela (2000b, p. 210) makes a similar observation: "To be sure agreement making can well change more in the game than has been assumed above: Due to its institutional and *quasi*-moral character it can change the payoffs of defection in a way that changes the whole nature of the game."

group members act according to plan. To implement this in a simple way, we consider the update of the deontic ideality function to be uniform for all agents in the game.[16] Accordingly, updating a game with a group plan amounts to this: if an action profile is forbidden by the plan, its deontic ideality becomes 0; and if an action profile is permitted by the plan, its deontic ideality remains unchanged. Or equivalently, an action profile a is deontically ideal in $S{\uparrow}P_{\mathcal{G}}$ if and only if (1) the action profile a is deontically ideal in S; and (2) the group's component action $a_{\mathcal{G}}$ of the action profile a is a group action in $P_{\mathcal{G}}$:

Definition 2.5 (Plan Updates)**.** *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Then the update of S with $P_{\mathcal{G}}$, notation: $S{\uparrow}P_{\mathcal{G}}$, is the deontic game $S{\uparrow}P_{\mathcal{G}} = \langle N, (A_i), d^{\uparrow}\rangle$, where*

$$d^{\uparrow}(a) \quad = \quad \begin{cases} d(a), & \text{if } a_{\mathcal{G}} \in P_{\mathcal{G}} \\ 0, & \text{if } a_{\mathcal{G}} \notin P_{\mathcal{G}}. \end{cases}$$

Because an update of a deontic game with a group plan might change the deontic ideality of the game's action profiles, it might also change the dominance ordering of any group's group actions. There is one exception. Updating a deontic game with a group plan leaves the dominance ordering of the group actions in the plan unaffected (all claims are proved in Appendix B):

Observation 2.1. *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Let $a_{\mathcal{G}}, a^*_{\mathcal{G}} \in P_{\mathcal{G}}$. Then*

$$a_{\mathcal{G}} \geq_S a^*_{\mathcal{G}} \quad \textit{iff} \quad a_{\mathcal{G}} \geq_{S{\uparrow}P_{\mathcal{G}}} a^*_{\mathcal{G}}.$$

Although an update with a group plan does not change the dominance ordering of the group actions in the plan, it might change the dominance ordering of the individual actions that are component actions of the group actions in the plan.

[16] These idealizing assumptions are similar to those that underlie the logic of public announcements (Plaza, 1989). They were relaxed in the logic of epistemic actions (Baltag et al., 1998). Van Ditmarsch et al. (2007, Chs. 4 and 5) provide textbook presentations of these logics.

Ideally, an update with a group plan should change the dominance ordering of these individual actions in such a way that an individual action is a component of one of the group actions in the plan if and only if that individual action is an admissible individual action in the deontic game that results from updating the original deontic game with the plan. Not all group plans have this updating property. We show that if a group plan is a good plan (in the sense defined below), then it has this updating property.

2.4 Good Plans and Bad Plans

There are good plans and bad plans. A good plan guarantees that if every group member fulfils her member obligation specified by the plan, then the group itself fulfils its collective obligation.[17] Or equivalently, a good plan guarantees that if the group itself does *not* fulfil its collective obligation, then at least one group member does *not* fulfil her member obligation specified by the plan.[18] A bad plan fails to do so. By agreeing on a bad plan, even though it thereby becomes common knowledge that the plan has been adopted, the group members do not rule out the possibility that every group member fulfils her member obligation specified by the plan, while the composition of the group members' actions does not amount to a group action that fulfils their collective obligation. Accordingly, there are two conditions on fulfilling a collective obligation by way of a group plan: first, the group members must act according to plan, and second, the plan itself must be good. But what makes a group plan a good plan? We submit that a plan is a good plan if and only if it is *optimal* and *interchangeable*.

[17] Compare Regan (1980, p. 138 – notation adapted): "If the members of $\mathcal{G}$ all do their part in the best pattern of behaviour for the members of $\mathcal{G}$ given the behaviour of non-members, it is clear that the members of $\mathcal{G}$ produce the best consequences possible as a group." We argue below that the truth of this conditional depends on the structure of the plan: if the plan is what we call 'optimal' and 'interchangeable', then the conditional is true.

[18] Bratman (2014, p. 34) claims that in basic cases "violations of these norms of social rationality will be constituted by violations, by one or more participants, of associated norms of individual planning agency". See also Jackson (1987, p. 107), who agrees with us that the conditional "If a group act is wrong, at least one of its constituent individual acts is wrong" is not a truth of logic.

2.4.1 Optimal Plans

The aim of a group plan is to fulfil a collective obligation. A group fulfils its collective obligation if and only if it performs an admissible group action. A good plan should hence be designed such that if all group members act according to plan, then the composition of their actions amounts to an admissible group action. The first design requirement of a good plan therefore is that the plan only consists of admissible group actions (as we shall see below, this is a necessary but not a sufficient condition). We say that a group plan is *optimal* if and only if it is a non-empty subset of the group's admissible group actions:

Definition 2.6 (Optimal Plans)**.** *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Then $P_{\mathcal{G}}$ is* optimal *if and only if $P_{\mathcal{G}} \neq \emptyset$ and $P_{\mathcal{G}} \subseteq Admissible_S(\mathcal{G})$.*

What happens if we update a deontic game with such an optimal plan? Ruling out flat deontic games (in which there is no deontically ideal action profile), an update of a deontic game with an optimal plan does not introduce admissible group actions that were not in the plan and does not eliminate admissible group actions that were in the plan. The set of admissible group actions in the deontic game that results from updating a deontic game with an optimal plan equals the set of group actions in the plan:

Observation 2.2. *Let $S = \langle N, (A_i), d\rangle$ be a non-flat deontic game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$ be optimal. Then*

$$Admissible_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{G}) = P_{\mathcal{G}}.$$

This observation implies that every group action that is admissible in the deontic game $S{\uparrow}P_{\mathcal{G}}$ that results from updating a non-flat deontic game S with an optimal plan $P_{\mathcal{G}}$ is an admissible group action in the original deontic game S, that is, $Admissible_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{G}) \subseteq Admissible_S(\mathcal{G})$. Accordingly, an update with an optimal plan highlights particular admissible group actions in the original deontic game.

However, it is not sufficient to require of a good plan that it be optimal. Indeed, if a plan consists of several admissible group actions and if each group member does her share of an admissible group action in the plan, then the composition of the group members' actions need not be an admissible group action. Writing on action profiles in games with individual utility functions rather than on group actions in games with a deontic ideality function, David Gauthier (1975, p. 201) observes: "If each person performs an action which has a best equilibrium as a possible outcome, and if there are several best equilibria, then the outcome need not be a best equilibrium." Because of this, coordination theorists standardly assume that a solution to a coordination problem consists in singling out a unique action profile:

> Several best equilibria are too many of a good thing. [...] What we need is a way to restructure our conception of the situation so that we are left with but one. We must restrict the possible actions which we consider, in such a way that we convert our representation of the situation into one with but one best equilibrium. (Gauthier, 1975, p. 210)[19]

In the present setting, we do not need this methodological assumption. To tackle our current problem of specifying the conditions under which a group plan successfully coordinates the actions of group members, we do not have to assume that the agents must somehow agree on a unique group action. We show that it suffices to require of good plans not only that they be optimal, but also interchangeable.

[19]Compare Harsanyi and Selten (1988, p. 13): "Clearly a theory telling us no more than that the outcome can be any one of these equilibrium points will not give us much useful information. We need a theory selecting one equilibrium point as the solution of the game."

2.4.2 Interchangeable Plans

An optimal plan does not guarantee that if all group members act according to plan, then the group itself performs an admissible group action. We therefore need an additional requirement. This second design requirement of a good plan is that the plan be closed under component individual actions. We say that a group plan is *interchangeable* if and only if for any two group actions in the plan it holds that the composition of a group member's contribution to the first group action and the joint contribution of all other group members to the second group action amounts to a group action that is also in the plan:[20]

Definition 2.7 (Interchangeable Plans)**.** *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Then $P_{\mathcal{G}}$ is* interchangeable *if and only if for all $a_{\mathcal{G}}, a'_{\mathcal{G}} \in P_{\mathcal{G}}$ and all $i \in \mathcal{G}$ it holds that $(a_i, a'_{\mathcal{G}-i}) \in P_{\mathcal{G}}$.*[21]

A few examples: in deontic game S_3 of Figure 2.3 the plan $\{(a_i, a_j)\}$ is interchangeable (as is every plan consisting of a single group action). Likewise, $\{(a_i, a_j), (a_i, a'_j), (a_i, a''_j)\}$ is interchangeable. The plan $\{(a_i, a_j), (a_i, a'_j), (a'_i, a_j)\}$ is not, because (a'_i, a'_j) is not in it. Finally, $\{(a_i, a_j), (a_i, a''_j), (a'_i, a_j), (a'_i, a''_j)\}$ is interchangeable.

	a_j	a'_j	a''_j
a_i	1	1	0
a'_i	0	0	1
a''_i	0	1	0

Figure 2.3: *Deontic game S_3.*

[20] The concept of interchangeability goes back to Nash (1951, p. 290).
[21] Observation 3.3 in Chapter 3 provides a logical characterization of interchangeability.

Our definition implies that a plan is interchangeable if and only if for any two group actions in the plan it holds that the composition of a subgroup's contribution to the first group action and the joint contribution of all other group members to the second group action amounts to a group action that is also in the plan:

Observation 2.3. *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Then $P_{\mathcal{G}}$ is interchangeable if and only if for all $a_{\mathcal{G}}, a'_{\mathcal{G}} \in P_{\mathcal{G}}$ and all $\mathcal{F} \subseteq \mathcal{G}$ it holds that $(a_{\mathcal{F}}, a'_{\mathcal{G}-\mathcal{F}}) \in P_{\mathcal{G}}$.*

We have argued above that a group plan does not need to single out a unique group action to successfully coordinate the individual actions of the group members. To make this clear, consider again the deontic game S_3 of Figure 2.3. To solve their coordination problem, group members i and j must agree on a plan that successfully coordinates their actions. Even though it consists of several group actions, the optimal and interchangeable plan $\{(a_i, a_j), (a_i, a'_j)\}$ does the job. The plan specifies the following member obligations on agents i and j: agent i ought to perform action a_i and agent j ought to perform one of the actions a_j and a'_j. The plan hence gives some leeway to agent j. Nonetheless, the plan ensures that if both group members act according to plan, they perform an admissible group action and hence the group itself fulfils its collective obligation.

This point can be made fully general: an optimal and interchangeable plan *guarantees* that if every group member fulfils her member obligation specified by the plan, then the group itself fulfils its collective obligation. Let us see why this is so. If every group member fulfils her member obligation specified by an optimal and interchangeable plan, then every group member performs an action that is a component action of one of the group actions in the plan. Because the plan is interchangeable, the combination of the actions performed by the group members must also be in the plan. Because the plan is optimal, it only contains admissible group actions, and hence this combination of group member actions is an admissible group action. The group members hence perform an admissible

group action and thus fulfil the group's collective obligation. Therefore, if every group member acts according to an optimal and interchangeable plan, then the group itself fulfils its collective obligation.

2.4.3 Updates with Optimal and Interchangeable Plans

Our central theorem concerns what happens if we update a deontic game with an optimal and interchangeable group plan. We show that if we update a non-flat deontic game with such a plan, then for any subgroup of this group it holds that an action of that subgroup is an admissible subgroup action in the deontic game that results from updating the original deontic game with the plan if and only if that subgroup action is a component subgroup action of one of the group actions in the plan:

Theorem 2.1. *Let $S = \langle N, (A_i), d\rangle$ be a non-flat deontic game. Let $\mathcal{F} \subseteq \mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$ be optimal and interchangeable. Then*

$$Admissible_{S\uparrow P_{\mathcal{G}}}(\mathcal{F}) = \{a_{\mathcal{F}} : a_{\mathcal{G}} \in P_{\mathcal{G}}\}.$$

Our theorem shows that updates of deontic games with optimal and interchangeable plans relate collective obligations to individual obligations. Indeed, by setting $\mathcal{F} = \{i\}$ for any individual member i of $\mathcal{G}$ we obtain $Admissible_{S\uparrow P_{\mathcal{G}}}(i) = \{a_i : a_{\mathcal{G}} \in P_{\mathcal{G}}\}$. Accordingly, if we update a non-flat deontic game with an optimal and interchangeable group plan, then for every group member it holds that she fulfils her member obligation specified by the plan if and only if she fulfils her individual obligation in the deontic game that results from updating the original deontic game with the plan. Consequently, under the given assumptions, a group member acts according to plan if and only if she performs an admissible individual action in the changed decision context. It follows that, under the given

assumptions, if every group member performs an admissible individual action in the changed decision context, then the group itself performs an admissible group action.

If we think of acting rationally as performing those actions that best promote deontic ideality, then we could say that an individual action is *individually rational* if and only if it is an admissible individual action, and that a group action is *collectively rational* if and only if it is an admissible group action. Within the narrow confines of our idealizing assumptions, Theorem 2.1 would then relate collective rationality to individual rationality and thereby give a tentative answer to Bratman's questions:

> Are there norms that are in some way fundamental to shared agency? If so, how precisely are they related to shared agency, and how are they related to norms of individual rationality? (Bratman, 2014, p. 4)

The answer runs as follows. A group action is collectively rational if and only if it is an admissible group action. To perform an admissible group action, the group members must adopt a good plan to ensure that the combination of their actions is an admissible group action. After the public adoption of such a plan, a group member's action is rational with respect to the group's objective of performing an admissible group action if and only if this action is permitted by the plan.[22] Moreover, when the plan has been publicly adopted and thereby changed the context in which agents make a decision about what to do, an individual action is individually rational if and only if it is an admissible individual action in the changed decision context. Our theorem shows that, under the given assumptions, a group member's action is rational with respect to the group's objective of performing an admissible group action if and only if that action is

[22]Compare (Gold, 2012, p. 185): "The basic idea is that, when an individual reasons as a member of a team, she considers which *combination* of actions by members of the team would best promote the team's objective, and then performs her part of that combination. The rationality of each individual's action derives from the rationality of the joint action of the team." Sugden (1993, p. 87), Anderson (2001, pp. 28–30), and Bacharach (2006, p. 136) defend similar notions of group member rationality.

individually rational in the decision context that results from the public adoption of the plan. It follows that, under the given assumptions, if every group member performs an individually rational action in the changed decision context, then the group itself performs a collectively rational group action. Therefore, if we think of acting rationally as performing those actions that best promote deontic ideality, our theorem would establish a connection between collective rationality and individual rationality.

2.5 Zero-Preserving Cooperation Games

Thus far, we have used deontic games involving a single deontic ideality function to study the relation between collective obligations, group plans, and individual actions. In this section we show that almost all our findings on deontic games transfer to 'cooperation games' involving group-relative utility functions, provided these utility functions are 'zero-preserving'.[23] The only exception is the right-to-left inclusion of Theorem 2.1. In the context of zero-preserving cooperation games this property is undesirable. To show all of this let us first introduce cooperation games.

A cooperation game includes a game form and a non-negative group utility for each group of agents:

Definition 2.8 (Cooperation Game)**.** *A* cooperation game *S is a triple $\langle N, (A_i), (u_{\mathcal{H}})\rangle$, where $\langle N, (A_i)\rangle$ is a game form, and for each group of agents $\mathcal{H}$ from N it holds that $u_{\mathcal{H}}$ is a utility function that assigns to each action profile a in A $(= \times_{i\in N} A_i)$ a utility $u_{\mathcal{H}}(a) \in \mathbb{R}_{\geq 0}$.*

To prove our claims about cooperation games we rule out cooperation games in which a group assigns to every action profile a utility of 0. Such games are *flat with respect to that group*:

[23] For clarity's sake, we also provide definitions similar to 2.1, 2.2, 2.3, 2.4 for cooperation games.

Definition 2.9 ($\mathcal{G}$-Flat)**.** *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a cooperation game. Let $\mathcal{G} \subseteq N$. Then S is* $\mathcal{G}$-flat *if and only if for every action profile $a \in A$ it holds that $u_{\mathcal{G}}(a) = 0$.*

To define the $\mathcal{G}$-admissible group actions that are available to a group $\mathcal{G}$, we first define a dominance ordering: group action $a_{\mathcal{G}}$ *weakly $\mathcal{G}$-dominates* group action $a'_{\mathcal{G}}$ in cooperation game S (notation: $a_{\mathcal{G}} \succeq_S a'_{\mathcal{G}}$) if and only if $a_{\mathcal{G}}$ promotes the utility of group $\mathcal{G}$ in S at least as well as $a'_{\mathcal{G}}$, regardless of what $\mathcal{G}$'s non-members do:

Definition 2.10 ($\mathcal{G}$-Dominance)**.** *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a cooperation game. Let $\mathcal{G} \subseteq N$. Let $a_{\mathcal{G}}, a'_{\mathcal{G}} \in A_{\mathcal{G}}$. Then*

$$a_{\mathcal{G}} \succeq_S a'_{\mathcal{G}} \quad \textit{iff} \quad \textit{for all } a''_{-\mathcal{G}} \in A_{-\mathcal{G}} \textit{ it holds that } u_{\mathcal{G}}(a_{\mathcal{G}}, a''_{-\mathcal{G}}) \geq u_{\mathcal{G}}(a'_{\mathcal{G}}, a''_{-\mathcal{G}}).$$

Strong $\mathcal{G}$-dominance is defined in terms of weak $\mathcal{G}$-dominance: $a_{\mathcal{G}} \succ_S a'_{\mathcal{G}}$ if and only if $a_{\mathcal{G}} \succeq_S a'_{\mathcal{G}}$ and $a'_{\mathcal{G}} \not\succeq_S a_{\mathcal{G}}$.

A group action is $\mathcal{G}$-admissible in cooperation game S if and only if it is not strongly $\mathcal{G}$-dominated by any group action that is available to the group in S:

Definition 2.11 ($\mathcal{G}$-Admissible Actions)**.** *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a cooperation game. Let $\mathcal{G} \subseteq N$. Then*

$$\textit{Admissible}_S(\mathcal{G}) \quad = \quad \{a_{\mathcal{G}} \in A_{\mathcal{G}} : \textit{ there is no } a'_{\mathcal{G}} \in A_{\mathcal{G}} \textit{ such that } a'_{\mathcal{G}} \succ_S a_{\mathcal{G}}\}.$$

Just as in deontic games, a group $\mathcal{G}$ fulfils its collective obligation in cooperation game S if and only if it performs one of its $\mathcal{G}$-admissible actions. Likewise, an agent i fulfils her individual obligation in cooperation game S if and only if she performs one of her i-admissible actions.

As before, a group plan $P_{\mathcal{G}}$ of group $\mathcal{G}$ in a cooperation game S can be any subset of the set $A_{\mathcal{G}}$ of group actions that are available to group $\mathcal{G}$ in S. Again, given an adopted group plan $P_{\mathcal{G}}$ and a group member i of $\mathcal{G}$, agent i fulfils her member obligation specified by the plan if and only if she performs an action that is her component action of one of the group actions in the plan.

In analogy to the plan updates of § 2.3.1, we now model the public adoption of a group plan in a cooperation game as an update of the game's utility functions. Updating a cooperation game with a group plan amounts to the following: if an action profile is forbidden by the plan, then for every group the group utility of that action profile becomes 0; and if an action profile is permitted by the plan, then for every group the group utility of that action profile remains unchanged:

Definition 2.12 (Plan Updates)**.** *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a cooperation game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Then the update of S with $P_{\mathcal{G}}$, notation: $S{\uparrow}P_{\mathcal{G}}$, is the cooperation game $S{\uparrow}P_{\mathcal{G}} = \langle N, (A_i), (u^{\uparrow}_{\mathcal{H}})\rangle$, where*

$$u^{\uparrow}_{\mathcal{H}}(a) \quad = \quad \begin{cases} u_{\mathcal{H}}(a), & \text{if } a_{\mathcal{G}} \in P_{\mathcal{G}} \\ 0, & \text{if } a_{\mathcal{G}} \notin P_{\mathcal{G}}. \end{cases}$$

As in deontic games, an update of a cooperation game with a group plan leaves the dominance ordering of the group actions in the plan unaffected:

Observation 2.4. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a cooperation game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Let $a_{\mathcal{G}}, a^*_{\mathcal{G}} \in P_{\mathcal{G}}$. Then*

$$a_{\mathcal{G}} \succeq_S a^*_{\mathcal{G}} \quad \text{iff} \quad a_{\mathcal{G}} \succeq_{S{\uparrow}P_{\mathcal{G}}} a^*_{\mathcal{G}}.$$

$\mathcal{G}$-optimal plans and interchangeable plans are formally defined as in Definitions 2.6 and 2.7 above. A group plan is *$\mathcal{G}$-optimal* if and only if it is a non-empty subset of the group's $\mathcal{G}$-admissible actions. A group plan is *interchangeable* if and only if for any two group actions in the plan it holds that the composition of a group member's contribution to the first group action and the joint contribution of all other group members to the second group action amounts to a group action that is also in the plan. Again, a $\mathcal{G}$-optimal and interchangeable plan guarantees that if every group member fulfils her member obligation specified by the plan, then the group itself fulfils its collective obligation. For an informal proof of this conditional, see the end of § 2.4.2.

After updating a non-$\mathcal{G}$-flat cooperation game with a $\mathcal{G}$-optimal group plan $P_{\mathcal{G}}$ it holds that a group action is a $\mathcal{G}$-admissible group action in the cooperation game that results from the update if and only if it is one of the group actions in the plan:

Observation 2.5. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a non-$\mathcal{G}$-flat cooperation game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$ be $\mathcal{G}$-optimal. Then*

$$Admissible_{S\uparrow P_{\mathcal{G}}}(\mathcal{G}) = P_{\mathcal{G}}.$$

To transfer our findings on deontic games to cooperation games, we must relate group utilities to subgroup utilities. To do this, we require the relation between subgroup utilities and group utilities to be *zero-preserving*. A cooperation game is zero-preserving if and only if there is no action profile in the game which has a non-zero utility for some group, while it does have zero utility for some subgroup of that group:

Definition 2.13 (Zero-Preservation)**.** *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a cooperation game. Then S is* zero-preserving *if and only if for all non-empty groups $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq N$ and for all action profiles $a \in A$ it holds that if $u_{\mathcal{F}_1}(a) = 0$, then $u_{\mathcal{F}_2}(a) = 0$.*

Although this might seem a rather strong condition, any cooperation game can be transformed into a zero-preserving one by linearly transforming the game's utility functions: adding 1 to each utility in the game makes it zero-preserving. Apart from using linear transformations to ensure that a cooperation game is zero-preserving, some social welfare functions always ensure zero-preservation. The Rawlsian social welfare function $u_{\mathcal{H}}(a) = \min_{i\in\mathcal{H}}(u_i(a))$ is but one example. Other social welfare functions, such as summing or averaging, only meet this condition in some cooperation games – the Hi-Lo game that we discuss in § 2.5.1 is a case in point.[24]

Plan updates sustain zero-preservation:

[24] Team-reasoning theorists are not in favour of imposing conceptual constraints on social welfare functions. Because his theory of team reasoning does not solely concern cooperation for mutual

Observation 2.6. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a cooperation game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Then*

If S is zero-preserving, then $S{\uparrow}P_{\mathcal{G}}$ is zero-preserving.

Finally, the left-to-right inclusion of Theorem 2.1 also holds for cooperation games. We show that if we update a zero-preserving non-$\mathcal{G}$-flat cooperation game with a $\mathcal{G}$-optimal group plan $P_{\mathcal{G}}$, then for any subgroup $\mathcal{F}$ of $\mathcal{G}$ it holds that if an action of that subgroup is an $\mathcal{F}$-admissible subgroup action in the cooperation game that results from updating the original cooperation game with the plan, then that subgroup action is a component subgroup action of one of the group actions in the plan:

Observation 2.7. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a zero-preserving non-$\mathcal{G}$-flat cooperation game. Let $\mathcal{F} \subseteq \mathcal{G} \subseteq N$ be non-empty and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$ be $\mathcal{G}$-optimal. Then*

$Admissible_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{F}) \subseteq \{a_{\mathcal{F}} : a_{\mathcal{G}} \in P_{\mathcal{G}}\}$.

Analogous to Theorem 2.1, our observation shows also that updates of cooperation games with $\mathcal{G}$-optimal plans relate collective obligations to individual obligations. Indeed, by setting $\mathcal{F} = \{i\}$ for any individual member i of $\mathcal{G}$ we obtain $Admissible_{S{\uparrow}P_{\mathcal{G}}}(i) \subseteq \{a_i : a_{\mathcal{G}} \in P_{\mathcal{G}}\}$. Accordingly, if we update a zero-preserving non-$\mathcal{G}$-flat cooperation game with a $\mathcal{G}$-optimal plan, then for every group member it holds that she fulfils her member obligation specified by the plan if she fulfils her individual obligation in the cooperation game that results from updating the original cooperation game with the plan. Consequently, under the given assumptions, a group member i acts according to plan if she performs an i-admissible individual action in the changed decision context. (Notice that, until now, we have not required the plan to be interchangeable.) It follows that, under

advantage, Bacharach "allowed in principle that the group objective might be welfare decreasing for some members" (Gold, 2012, p. 195). According to Sugden (2000, p. 176), "the preferences of a team are not necessarily reducible to, or capable of being constructed out of, the preferences that govern the choices that the members of the team make as individuals."

the given assumptions, a $\mathcal{G}$-optimal and interchangeable plan guarantees that if each group member i performs an i-admissible individual action in the changed decision context, then the group itself performs a $\mathcal{G}$-admissible group action.

The right-to-left inclusion does not hold, because plan updates do not affect utilities of action profiles that are permitted by the plan. This means that if we update a zero-preserving non-$\mathcal{G}$-flat cooperation game with a $\mathcal{G}$-optimal and interchangeable group plan, then it might be that a group member acts according to plan without performing an i-admissible individual action in the changed decision context. To see this, consider the zero-preserving non-$\mathcal{G}$-flat cooperation game S_4 of Figure 2.4, where each action profile a is assigned a triple of utilities $\langle u_i(a), u_j(a), u_{i,j}(a)\rangle$ and $\mathcal{G} := \{i, j\}$. (The utility function of the empty group is suppressed.)

	a'_j	a''_j
a'_i	3, 4, 3	0, 1, 0
a''_i	4, 3, 3	2, 5, 2

Figure 2.4: *Cooperation game* S_4.

Updating S_4 with the $\mathcal{G}$-optimal and interchangeable plan $P_{i,j} = \{(a'_i, a'_j), (a''_i, a'_j)\}$ results in the cooperation game $S_4{\uparrow}P_{i,j}$ of Figure 2.5, where it holds that $a'_i \in \{a_i : a_{i,j} \in P_{i,j}\}$ and $a'_i \notin \mathit{Admissible}_{S_4{\uparrow}P_{i,j}}(i)$.

	a'_j	a''_j
a'_i	3, 4, 3	0, 0, 0
a''_i	4, 3, 3	0, 0, 0

Figure 2.5: *Cooperation game* $S_4{\uparrow}P_{i,j}$.

2.5.1 The Team-Reasoning Account of Cooperation

Our discussion of cooperation games permits us to compare our account of the relation between collective obligations, group plans, and individual actions and the team-reasoning account of cooperation. (See § 4.2.1 and § 5.4 for a more detailed introduction to the idea of team reasoning.) Disregarding the differences in their accounts of cooperation, Bacharach and Sugden both argue that if we wish to give a game-theoretical explanation of cooperation in single-shot games, then an individualist perspective will not do.[25] Therefore, they propose to extend traditional game theory with group notions. Given a group of agents, team reasoning describes the reasoning process by which a group member first identifies the best combination of actions that the group members can perform and then decides to perform the individual action that is her part in that combination. To see how this works, consider the Hi-Lo cooperation game S_5 of Figure 2.6, where each action profile a is assigned a triple of utilities $\langle u_i(a), u_j(a), u_{i,j}(a)\rangle$ such that $u_{i,j}(a)$ is the sum of $u_i(a)$ and $u_j(a)$:

	high	*low*
high	5, 5, 10	0, 0, 0
low	0, 0, 0	1, 1, 2

Figure 2.6: *Hi-Lo cooperation game S_5.*

Agent i is engaged in team reasoning only if she reaches her decision about what to do by reasoning from the standpoint of the team consisting of i and j. First, it will then be clear to her that, of the four courses of action that are available to the group, the combination (*high*, *high*) is the best. She therefore decides to do her part in that combination and perform action *high*. Reasoning in an analogous fashion, agent j decides to perform action *high*. Accordingly, if both agents engage in team

[25] Gold (2012) provides a detailed comparison of Bacharach's and Sugden's accounts of team reasoning.

reasoning, they will perform the best group action. In contrast to traditional game theory, the team-reasoning account explains cooperation in Hi-Lo games even if communication is impossible.

Contrary to the team-reasoning account of cooperation, we assume that communication is possible and that agreement is unproblematic. Our research agenda therefore differs from that set by Bacharach and Sugden. Our account of cooperation starts from the assumption that the agents i and j are able to communicate and make agreements about what to do. If they decide to act as a group, they should abandon their individual perspective and evaluate the actions available to them from the standpoint of the group.[26] Because it will be clear to them that $(high, high)$ is the only $\{i, j\}$-admissible group action, they should agree on the group plan $\{(high, high)\}$. At the moment they have publicly adopted this group plan, agent i has the member obligation to perform action *high* and agent j has the member obligation to perform *high*. Moreover, by publicly adopting this group plan, the decision context is changed into cooperation game $S_5{\uparrow}\{(high, high)\}$ of Figure 2.7. In this changed decision context, agent i has the individual obligation to perform action *high* and agent j has the individual obligation to perform action *high*.

	high	*low*
high	5, 5, 10	0, 0, 0
low	0, 0, 0	0, 0, 0

Figure 2.7: *Cooperation game* $S_5{\uparrow}\{(high, high)\}$.

Our analysis hence suggests that, given a decision context in which the agents are able to communicate and make agreements, it holds that if they decide to act as a group $\mathcal{G}$ of agents, and if there is a unique $\mathcal{G}$-admissible group action, then the group members should agree on the group plan that only consists of this unique

[26] Note that pre-play communication does not automatically lead to group identifying or team reasoning. Gold (2012, p. 200) observes: "For both Bacharach and Sugden there are potential gaps between noticing the group and group identifying, and between group identifying and team reasoning."

$\mathcal{G}$-admissible group action. The member obligations specified by the plan and the individual obligations in the decision context that results from publicly adopting the plan coincide with the recommendations that issue from the team-reasoning account. Therefore, if there is a unique $\mathcal{G}$-admissible group action, our account and the team-reasoning account of cooperation give similar recommendations to group members. Things are different, however, if there are several $\mathcal{G}$-admissible group actions: whereas the team-reasoning account has nothing much to offer, our analysis still applies.[27]

We have argued that if the group members design and adopt a group plan to solve their coordination problem, and if we allow a group plan to consist of *several* group actions (thereby giving some leeway to at least some group members), the plan might fail to guarantee that if every group member acts according to plan, then the combination of their actions is a $\mathcal{G}$-admissible group action. Some such plans do guarantee this, however. The focus of our study has therefore been on the structural conditions that a group plan must meet in order to successfully coordinate the individual actions of the group members, and on what happens to a decision context when such a group plan is publicly adopted.

2.6 Conclusion

What, if anything, does our account of collective obligations, group plans, and individual actions tell us about the concept of collective blameworthiness, that is, backward-looking collective moral responsibility? The complexity of this concept has led some to argue that formal methods are hardly any use in the study of collective moral responsibility. Margaret Gilbert, for example, writes:

> What does the blameworthiness of the collective's act imply about the personal blameworthiness of any one member of that collective? From

[27] In Chapter 5 I refine the reasoning method akin to team reasoning to deal with scenarios in which there are several $\mathcal{G}$-admissible group actions, and in Chapter 4 I study and refine the intentions that result from team reasoning.

> a logical point of view, the short answer is: *nothing*. Everything depends on the details of a given member's particular situation. (Gilbert, 2006b, p. 109)[28]

Gilbert notes that some members of a blameworthy collective might have authorized others to decide and that this may make a difference to the attribution of blame to those members. At present, our formal analysis is too coarse to capture these structural aspects of collective moral responsibility. Nonetheless, even our relatively simple formalization of the relation between collective obligations, group plans, and individual actions brings to the fore some central aspects of the logic of backward-looking collective responsibility.

We take it that the concept of backward-looking individual responsibility supports the claim that not fulfilling an obligation is a necessary condition for being individually blameworthy: if an individual agent is individually blameworthy, then she has failed to fulfil an obligation. The converse does not hold, because she might have a plausible excuse for not doing what she ought to do. Analogously, we submit that not fulfilling a collective obligation is a necessary, but not a sufficient condition for being collectively blameworthy. We thus obtain the following implication: if a group is collectively blameworthy, then it does not fulfil a collective obligation.

Our account of the relation between collective obligations, group plans, and individual actions brought to light that a good plan has two key characteristics. First, a good plan guarantees that if every group member fulfils her member obligation specified by the plan, then the group itself fulfils its collective obligation. Second, if the group members, aiming to fulfil a collective obligation, publicly adopt a good plan to coordinate their individual actions, then for any group member it holds that if she fulfils her individual obligation in the decision context that results from publicly adopting the plan, then she fulfils her member oblig-

[28] Likewise, Isaacs (2011, p. 24) argues that "claims about collective moral responsibility neither entail nor are derivable from claims about individual moral responsibility".

ation specified by the plan. We thus obtain a second implication (which holds for both deontic games and cooperation games): given that the group members, aiming to fulfil a collective obligation, publicly adopt a good plan to coordinate their individual actions, it holds that their failure to fulfil the collective obligation *logically implies* that at least one of them failed to fulfil her member obligation specified by the plan, which in turn entails that at least one group member failed to fulfil her individual obligation in the changed decision context.

Combining the two implications, we conclude: given that the group members, aiming to fulfil a collective obligation, publicly adopt a good plan to coordinate their actions, it holds that the collective blameworthiness of the group *logically implies* that at least one of the group members failed to act according to plan, which in turn entails that at least one group member failed to fulfil her individual obligation in changed decision context.[29]

From the perspective of the modes of acting that are connected to levels of culpability, this discussion is best related to the *causal* mode of acting. In other words, this discussion can be taken to focus on the group members' conduct rather than the mental states that accompany it. With regard to the fulfilment of obligations, the three perspectives – collective, individual, and group member – can be specified as follows: a group should perform a group action that fulfils its collective obligation; an individual agent should perform an individual action that fulfils its individual obligation; and, in case a group plan has been adopted, a group member should perform an individual action that fulfils its member obligation.

Let us turn to the question of whether responsibility voids could exist when we concentrate on the causal mode of acting. If we assume that a failure to fulfil an obligation entails blameworthiness, then there is no responsibility void when agreement is unproblematic. That is, under these assumptions, the group mem-

[29] Again, a group member's failure to fulfil her obligations does not entail that she is personally blameworthy – she may have had good reasons not to fulfil her obligations.

bers should adopt a *good* plan to regulate their group action. Or, conversely, if responsibility voids are to exist, agreement must be problematic or communication needs to be restricted. If, however, for some reason or other, the group members fail to agree on a group plan or if they agree on a bad plan, then it is not so clear how being collectively blameworthy relates to the failure of individuals to fulfil their obligations. In the subsequent chapters I show that there may still be a relation between collective blameworthiness and individual blameworthiness by taking additional factors into consideration, such as knowledge, intentions, and practical reasoning.

This page intentionally contains only this sentence.

Appendix B

Collective Obligations, Group Plans, and Individual Actions

B.1 Deontic Games: Proofs

Lemma B.1 (Horty 2001, p. 74)**.** *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ be a non-empty group of agents. Then*

$Admissible_S(\mathcal{G}) \neq \emptyset$.

Proof. Assume that $Admissible_S(\mathcal{G}) = \emptyset$. By definition of S, it holds that $A_\mathcal{G}$ is finite and non-empty. Then for all $a'_\mathcal{G} \in A_\mathcal{G}$ there is an $a_\mathcal{G} \in A_\mathcal{G}$ such that $a_\mathcal{G} \succ_S a'_\mathcal{G}$. Because $A_\mathcal{G}$ is finite, there must be a cycle $a^n_\mathcal{G} \succ_S a^{n-1}_\mathcal{G} \succ_S \ldots \succ_S a^2_\mathcal{G} \succ_S a^1_\mathcal{G} = a^n_\mathcal{G}$ with all members $a^k_\mathcal{G}$ in $A_\mathcal{G}$. Because $\succ_S$ is transitive, it holds that $a^n_\mathcal{G} \succ_S a^n_\mathcal{G}$, that is, $a^n_\mathcal{G} \succeq_S a^n_\mathcal{G}$ and $a^n_\mathcal{G} \not\succeq_S a^n_\mathcal{G}$. Contradiction. Therefore, $Admissible_S(\mathcal{G}) \neq \emptyset$. □

Lemma B.2 (Kooi and Tamminga 2008, p. 9)**.** *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{F} \subseteq \mathcal{G} \subseteq N$. Let $a_\mathcal{F}, a'_\mathcal{F} \in A_\mathcal{F}$ and let $a''_{\mathcal{G}-\mathcal{F}} \in A_{\mathcal{G}-\mathcal{F}}$. Then*

If $a_\mathcal{F} \succeq_S a'_\mathcal{F}$, then $(a_\mathcal{F}, a''_{\mathcal{G}-\mathcal{F}}) \succeq_S (a'_\mathcal{F}, a''_{\mathcal{G}-\mathcal{F}})$.

Proof. Assume $a_{\mathcal{F}} \succeq_S a'_{\mathcal{F}}$. Let $a''_{\mathcal{G}-\mathcal{F}} \in A_{\mathcal{G}-\mathcal{F}}$. Suppose $a'''_{-\mathcal{G}} \in A_{-\mathcal{G}}$. Then $(a''_{\mathcal{G}-\mathcal{F}}, a'''_{-\mathcal{G}}) \in A_{-\mathcal{F}}$. Hence, by our assumption it must be that $d(a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}, a'''_{-\mathcal{G}}) \geq d(a'_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}, a'''_{-\mathcal{G}})$. Therefore, for all $a'''_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $d(a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}, a'''_{-\mathcal{G}}) \geq d(a'_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}, a'''_{-\mathcal{G}})$. Hence, $(a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}) \succeq_S (a'_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}})$. □

Lemma B.3. *Let $S = \langle N, (A_i), d\rangle$ be a non-flat deontic game. Let $\mathcal{G} \subseteq N$. Then*

If $a_{\mathcal{G}} \in Admissible_S(\mathcal{G})$, then there is an $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ such that $d(a_{\mathcal{G}}, a'_{-\mathcal{G}}) = 1$.

Proof. Suppose that $a_{\mathcal{G}} \in Admissible_S(\mathcal{G})$. Suppose that for every $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $d(a_{\mathcal{G}}, a'_{-\mathcal{G}}) = 0$. Because S is non-flat, there is an $a^* \in A$ such that $d(a^*) = 1$. Consider $a^*_{\mathcal{G}}$. It holds that $a^*_{\mathcal{G}} \succ_S a_{\mathcal{G}}$, that is (1) $a^*_{\mathcal{G}} \succeq_S a_{\mathcal{G}}$ and (2) $a_{\mathcal{G}} \not\succeq_S a^*_{\mathcal{G}}$. We have (1), because for every $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $d(a_{\mathcal{G}}, a'_{-\mathcal{G}}) = 0$. We have (2), because $d(a_{\mathcal{G}}, a^*_{-\mathcal{G}}) = 0$ and $d(a^*_{\mathcal{G}}, a^*_{-\mathcal{G}}) = d(a^*) = 1$. Hence, $a_{\mathcal{G}} \notin Admissible_S(\mathcal{G})$. Contradiction. □

Observation 2.1. *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Let $a_{\mathcal{G}}, a^*_{\mathcal{G}} \in P_{\mathcal{G}}$. Then*

$a_{\mathcal{G}} \succeq_S a^*_{\mathcal{G}}$ *iff* $a_{\mathcal{G}} \succeq_{S\uparrow P_{\mathcal{G}}} a^*_{\mathcal{G}}$.

Proof. It is given that $a_{\mathcal{G}}, a^*_{\mathcal{G}} \in P_{\mathcal{G}}$. By construction of $S\uparrow P_{\mathcal{G}}$, for every $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $d(a_{\mathcal{G}}, a'_{-\mathcal{G}}) = d^{\uparrow}(a_{\mathcal{G}}, a'_{-\mathcal{G}})$ and $d(a^*_{\mathcal{G}}, a'_{-\mathcal{G}}) = d^{\uparrow}(a^*_{\mathcal{G}}, a'_{-\mathcal{G}})$. Then the following four statements are equivalent: (1) $a_{\mathcal{G}} \succeq_S a^*_{\mathcal{G}}$; (2) for every $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $d(a_{\mathcal{G}}, a'_{-\mathcal{G}}) \geq d(a^*_{\mathcal{G}}, a'_{-\mathcal{G}})$; (3) for every $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $d^{\uparrow}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) \geq d^{\uparrow}(a^*_{\mathcal{G}}, a'_{-\mathcal{G}})$; and (4) $a_{\mathcal{G}} \succeq_{S\uparrow P_{\mathcal{G}}} a^*_{\mathcal{G}}$. □

Lemma B.4. *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}}$ be optimal. Then*

If S is non-flat, then $S\uparrow P_{\mathcal{G}}$ is non-flat.

Proof. Suppose that S is non-flat. Because $P_{\mathcal{G}}$ is optimal, there is an $a_{\mathcal{G}} \in P_{\mathcal{G}}$ such that $a_{\mathcal{G}} \in \mathit{Admissible}_S(\mathcal{G})$. By Lemma B.3 there is an $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ such that $d(a_{\mathcal{G}}, a'_{-\mathcal{G}}) = 1$. By the construction of $S{\uparrow}P_{\mathcal{G}}$, it must be that $d^{\uparrow}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) = 1$. Therefore, $S{\uparrow}P_{\mathcal{G}}$ is non-flat. □

Observation 2.2. *Let $S = \langle N, (A_i), d\rangle$ be a non-flat deontic game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$ be optimal. Then*

$$\mathit{Admissible}_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{G}) = P_{\mathcal{G}}.$$

Proof. (⊆) Suppose that $a_{\mathcal{G}} \in \mathit{Admissible}_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{G})$. By Lemma B.4, $S{\uparrow}P_{\mathcal{G}}$ is non-flat. By Lemma B.3, it must be that there is an $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ such that $d^{\uparrow}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) = 1$. By construction of $S{\uparrow}P_{\mathcal{G}}$, it must be that $a_{\mathcal{G}} \in P_{\mathcal{G}}$.

(⊇) Suppose that $a_{\mathcal{G}} \in P_{\mathcal{G}}$. It is given that $a_{\mathcal{G}} \in \mathit{Admissible}_S(\mathcal{G})$ and that S is non-flat. Suppose that $a_{\mathcal{G}} \notin \mathit{Admissible}_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{G})$. Then there is an $a^*_{\mathcal{G}} \in \mathit{Admissible}_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{G})$ such that $a^*_{\mathcal{G}} \succ_{S{\uparrow}P_{\mathcal{G}}} a_{\mathcal{G}}$. By the first part of this proof, $a^*_{\mathcal{G}} \in P_{\mathcal{G}}$. By Observation 2.1, it must be that $a^*_{\mathcal{G}} \succ_S a_{\mathcal{G}}$. Hence, $a_{\mathcal{G}} \notin \mathit{Admissible}_S(\mathcal{G})$. Contradiction. Therefore, $a_{\mathcal{G}} \in \mathit{Admissible}_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{G})$. □

Observation 2.3. *Let $S = \langle N, (A_i), d\rangle$ be a deontic game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Then $P_{\mathcal{G}}$ is interchangeable if and only if for all $a_{\mathcal{G}}, a'_{\mathcal{G}} \in P_{\mathcal{G}}$ and all $\mathcal{F} \subseteq \mathcal{G}$ it holds that $(a_{\mathcal{F}}, a'_{\mathcal{G}-\mathcal{F}}) \in P_{\mathcal{G}}$.*

Proof. By straightforward induction on the number of members of $\mathcal{F}$. □

Theorem 2.1. *Let $S = \langle N, (A_i), d\rangle$ be a non-flat deontic game. Let $\mathcal{F} \subseteq \mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$ be optimal and interchangeable. Then*

$$\mathit{Admissible}_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{F}) = \{a_{\mathcal{F}} : a_{\mathcal{G}} \in P_{\mathcal{G}}\}.$$

Proof. (⊆) Suppose that $a_{\mathcal{F}} \in \mathit{Admissible}_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{F})$. By Lemma B.3, there is an $a'_{-\mathcal{F}} \in A_{-\mathcal{F}}$ such that $d^{\uparrow}(a_{\mathcal{F}}, a'_{-\mathcal{F}}) = 1$. By the construction of $S{\uparrow}P_{\mathcal{G}}$, it must be that $(a_{\mathcal{F}}, a'_{\mathcal{G}-\mathcal{F}}) \in P_{\mathcal{G}}$. Therefore, $a_{\mathcal{F}} \in \{a_{\mathcal{F}} : a_{\mathcal{G}} \in P_{\mathcal{G}}\}$.

($\supseteq$) Suppose that $a_{\mathcal{F}} \in \{a_{\mathcal{F}} : a_{\mathcal{G}} \in P_{\mathcal{G}}\}$. Then there is an $a'_{\mathcal{G}-\mathcal{F}} \in A_{\mathcal{G}-\mathcal{F}}$ such that $(a_{\mathcal{F}}, a'_{\mathcal{G}-\mathcal{F}}) \in P_{\mathcal{G}}$. Suppose that $a_{\mathcal{F}} \notin \mathit{Admissible}_{S\uparrow P_{\mathcal{G}}}(\mathcal{F})$. Then there is an $a^*_{\mathcal{F}} \in \mathit{Admissible}_{S\uparrow P_{\mathcal{G}}}(\mathcal{F})$ such that $a^*_{\mathcal{F}} \succ_{S\uparrow P_{\mathcal{G}}} a_{\mathcal{F}}$. Then there must be an $a^{**}_{-\mathcal{F}} \in A_{-\mathcal{F}}$ such that $d^{\uparrow}(a^*_{\mathcal{F}}, a^{**}_{-\mathcal{F}}) > d^{\uparrow}(a_{\mathcal{F}}, a^{**}_{-\mathcal{F}})$. Let $a'' = (a^*_{\mathcal{F}}, a^{**}_{-\mathcal{F}})$. Then $d^{\uparrow}(a'') = 1$. By construction of $S\uparrow P_{\mathcal{G}}$, it must be that $a''_{\mathcal{G}} = (a''_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}) = (a^*_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}) \in P_{\mathcal{G}}$. By Observation 2.3, it must be that $(a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}) \in P_{\mathcal{G}}$. Because $a^*_{\mathcal{F}} \succeq_{S\uparrow P_{\mathcal{G}}} a_{\mathcal{F}}$ and Lemma B.2, it must be that $(a^*_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}) \succeq_{S\uparrow P_{\mathcal{G}}} (a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}})$. Because it holds that $d^{\uparrow}(a^*_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}, a''_{-\mathcal{G}}) = d^{\uparrow}(a^*_{\mathcal{F}}, a^{**}_{-\mathcal{F}}) = 1$ and $d^{\uparrow}(a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}, a''_{-\mathcal{G}}) = d^{\uparrow}(a_{\mathcal{F}}, a^{**}_{-\mathcal{F}}) = 0$, we have $(a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}) \not\succeq_{S\uparrow P_{\mathcal{G}}} (a^*_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}})$. Then it must be that $(a^*_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}) \succ_{S\uparrow P_{\mathcal{G}}} (a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}})$. By Observation 2.1, it must be that $(a^*_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}) \succ_S (a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}})$. Then $(a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}) \notin \mathit{Admissible}_S(\mathcal{G})$ which contradicts the fact that $(a_{\mathcal{F}}, a''_{\mathcal{G}-\mathcal{F}}) \in P_{\mathcal{G}}$. Therefore, $a_{\mathcal{F}} \in \mathit{Admissible}_{S\uparrow P_{\mathcal{G}}}(\mathcal{F})$. □

B.2 Cooperation Games: Proofs

Lemma B.5. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a non-$\mathcal{G}$-flat cooperation game. Let $\mathcal{G} \subseteq N$. Then*

If $a_{\mathcal{G}} \in \mathit{Admissible}_S(\mathcal{G})$, then there is an $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ such that $u_{\mathcal{G}}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) > 0$.

Proof. Suppose that $a_{\mathcal{G}} \in \mathit{Admissible}_S(\mathcal{G})$. Suppose that for every $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $u_{\mathcal{G}}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) = 0$. Because S is non-$\mathcal{G}$-flat, there is an $a^* \in A$ such that $u_{\mathcal{G}}(a^*) > 0$. Consider $a^*_{\mathcal{G}}$. It holds that $a^*_{\mathcal{G}} \succ_S a_{\mathcal{G}}$, that is (1) $a^*_{\mathcal{G}} \succeq_S a_{\mathcal{G}}$ and (2) $a_{\mathcal{G}} \not\succeq_S a^*_{\mathcal{G}}$. We have (1), because for every $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $u_{\mathcal{G}}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) = 0$. We have (2), because $u_{\mathcal{G}}(a_{\mathcal{G}}, a^*_{-\mathcal{G}}) = 0$ and $u_{\mathcal{G}}(a^*_{\mathcal{G}}, a^*_{-\mathcal{G}}) = u_{\mathcal{G}}(a^*) > 0$. Hence, $a_{\mathcal{G}} \notin \mathit{Admissible}_S(\mathcal{G})$. Contradiction. □

Observation 2.4. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a cooperation game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Let $a_{\mathcal{G}}, a^*_{\mathcal{G}} \in P_{\mathcal{G}}$. Then*

$a_{\mathcal{G}} \succeq_S a^*_{\mathcal{G}}$ *iff* $a_{\mathcal{G}} \succeq_{S\uparrow P_{\mathcal{G}}} a^*_{\mathcal{G}}$.

Proof. It is given that $a_{\mathcal{G}}, a^*_{\mathcal{G}} \in P_{\mathcal{G}}$. By construction of $S\uparrow P_{\mathcal{G}}$, for every $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $u_{\mathcal{G}}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) = u^{\uparrow}_{\mathcal{G}}(a_{\mathcal{G}}, a'_{-\mathcal{G}})$ and $u_{\mathcal{G}}(a^*_{\mathcal{G}}, a'_{-\mathcal{G}}) = u^{\uparrow}_{\mathcal{G}}(a^*_{\mathcal{G}}, a'_{-\mathcal{G}})$. Then

the following four statements are equivalent: (1) $a_{\mathcal{G}} \succeq_S a^*_{\mathcal{G}}$; (2) for every $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $u_{\mathcal{G}}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) \geq u_{\mathcal{G}}(a^*_{\mathcal{G}}, a'_{-\mathcal{G}})$; (3) for every $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ it holds that $u^{\uparrow}_{\mathcal{G}}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) \geq u^{\uparrow}_{\mathcal{G}}(a^*_{\mathcal{G}}, a'_{-\mathcal{G}})$; and (4) $a_{\mathcal{G}} \succeq_{S\uparrow P_{\mathcal{G}}} a^*_{\mathcal{G}}$. □

Lemma B.6. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a cooperation game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$ be $\mathcal{G}$-optimal. Then*

If S is non-$\mathcal{G}$-flat, then $S\uparrow P_{\mathcal{G}}$ is non-$\mathcal{G}$-flat.

Proof. Suppose that S is non-$\mathcal{G}$-flat. Because $P_{\mathcal{G}}$ is $\mathcal{G}$-optimal, there is an $a_{\mathcal{G}} \in P_{\mathcal{G}}$ such that $a_{\mathcal{G}} \in \mathit{Admissible}_S(\mathcal{G})$. By Lemma B.5, there is an $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ such that $u_{\mathcal{G}}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) > 0$. By the construction of $S\uparrow P_{\mathcal{G}}$, it must be that $u^{\uparrow}_{\mathcal{G}}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) > 0$. Therefore, $S\uparrow P_{\mathcal{G}}$ is non-$\mathcal{G}$-flat. □

Observation 2.5. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a non-$\mathcal{G}$-flat cooperation game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$ be $\mathcal{G}$-optimal. Then*

$\mathit{Admissible}_{S\uparrow P_{\mathcal{G}}}(\mathcal{G}) = P_{\mathcal{G}}$.

Proof. (⊆) Suppose that $a_{\mathcal{G}} \in \mathit{Admissible}_{S\uparrow P_{\mathcal{G}}}(\mathcal{G})$. By Lemma B.6, $S\uparrow P_{\mathcal{G}}$ is non-$\mathcal{G}$-flat. By Lemma B.5, there is an $a'_{-\mathcal{G}} \in A_{-\mathcal{G}}$ such that $u^{\uparrow}_{\mathcal{G}}(a_{\mathcal{G}}, a'_{-\mathcal{G}}) > 0$. By construction of $S\uparrow P_{\mathcal{G}}$, it must be that $a_{\mathcal{G}} \in P_{\mathcal{G}}$.

(⊇) Suppose that $a_{\mathcal{G}} \in P_{\mathcal{G}}$. It is given that $a_{\mathcal{G}} \in \mathit{Admissible}_S(\mathcal{G})$ and that S is non-$\mathcal{G}$-flat. Suppose that $a_{\mathcal{G}} \notin \mathit{Admissible}_{S\uparrow P_{\mathcal{G}}}(\mathcal{G})$. Then there is an $a^*_{\mathcal{G}} \in \mathit{Admissible}_{S\uparrow P_{\mathcal{G}}}(\mathcal{G})$ such that $a^*_{\mathcal{G}} \succ_{S\uparrow P_{\mathcal{G}}} a_{\mathcal{G}}$. By the first part of this proof, $a^*_{\mathcal{G}} \in P_{\mathcal{G}}$. By Observation 2.4, it must be that $a^*_{\mathcal{G}} \succ_S a_{\mathcal{G}}$. Hence, $a_{\mathcal{G}} \notin \mathit{Admissible}_S(\mathcal{G})$. Contradiction. Therefore, $a_{\mathcal{G}} \in \mathit{Admissible}_{S\uparrow P_{\mathcal{G}}}(\mathcal{G})$. □

Observation 2.6. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a cooperation game. Let $\mathcal{G} \subseteq N$ and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$. Then*

If S is zero-preserving, then $S\uparrow P_{\mathcal{G}}$ is zero-preserving.

Proof. Suppose that S is zero-preserving. Suppose that $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq N$ are non-empty and suppose that $a \in A$. There are two cases: (1) $a_{\mathcal{G}} \in P_{\mathcal{G}}$ and (2) $a_{\mathcal{G}} \notin P_{\mathcal{G}}$. If we have (1), then by construction of $S{\uparrow}P_{\mathcal{G}}$ it holds that $u^{\uparrow}_{\mathcal{F}_1}(a) = u_{\mathcal{F}_1}(a)$ and $u^{\uparrow}_{\mathcal{F}_2}(a) = u_{\mathcal{F}_2}(a)$. It is given that if $u_{\mathcal{F}_1}(a) = 0$, then $u_{\mathcal{F}_2}(a) = 0$. Hence, if $u^{\uparrow}_{\mathcal{F}_1}(a) = 0$, then $u^{\uparrow}_{\mathcal{F}_2}(a) = 0$. If we have (2), then by construction of $S{\uparrow}P_{\mathcal{G}}$ it holds that $u^{\uparrow}_{\mathcal{F}_1}(a) = 0$ and $u^{\uparrow}_{\mathcal{F}_2}(a) = 0$. Hence, if $u^{\uparrow}_{\mathcal{F}_1}(a) = 0$, then $u^{\uparrow}_{\mathcal{F}_2}(a) = 0$. Therefore, $S{\uparrow}P_{\mathcal{G}}$ is zero-preserving. □

Lemma B.7. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a zero-preserving cooperation game. Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq N$ be non-empty. Then*

If S is $\mathcal{F}_1$-flat, then S is $\mathcal{F}_2$-flat.

Proof. Suppose that S is $\mathcal{F}_1$-flat. Then for every $a \in A$ it holds that $u_{\mathcal{F}_1}(a) = 0$. Because S is zero-preserving, for every $a \in A$ it holds that $u_{\mathcal{F}_2}(a) = 0$. Therefore, S is $\mathcal{F}_2$-flat. □

Observation 2.7. *Let $S = \langle N, (A_i), (u_{\mathcal{H}})\rangle$ be a zero-preserving non-$\mathcal{G}$-flat cooperation game. Let $\mathcal{F} \subseteq \mathcal{G} \subseteq N$ be non-empty and let $P_{\mathcal{G}} \subseteq A_{\mathcal{G}}$ be $\mathcal{G}$-optimal. Then*

$Admissible_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{F}) \subseteq \{a_{\mathcal{F}} : a_{\mathcal{G}} \in P_{\mathcal{G}}\}$.

Proof. Suppose that $a_{\mathcal{F}} \in Admissible_{S{\uparrow}P_{\mathcal{G}}}(\mathcal{F})$. By Lemma B.7, S is non-$\mathcal{F}$-flat. By Lemma B.5, there is an $a'_{-\mathcal{F}} \in A_{-\mathcal{F}}$ such that $u^{\uparrow}_{\mathcal{F}}(a_{\mathcal{F}}, a'_{-\mathcal{F}}) > 0$. By the construction of $S{\uparrow}P_{\mathcal{G}}$, it must be that $(a_{\mathcal{F}}, a'_{\mathcal{G}-\mathcal{F}}) \in P_{\mathcal{G}}$. Therefore, $a_{\mathcal{F}} \in \{a_{\mathcal{F}} : a_{\mathcal{G}} \in P_{\mathcal{G}}\}$. □

3

Collective Know-how

We want a computer program that decides what to do by inferring in a formal language that a certain strategy will achieve a certain goal. This requires formalizing concepts of causality, ability, and knowledge.

John McCarthy and Patrick Hayes (1969, p. 463)

3.1 Introduction

Our social lives are packed with collective endeavours without which it would be a lonely journey. An agent needs the capacity to reason about individual and collective abilities in order to coordinate her actions with companions in an optimal way. This holds both for human and for artificial agents. In this chapter I study the practical knowledge required for such collective activities. What knowledge do we possess when we know how to decorate a house together? Which epistemic conditions are sufficient or necessary to justify such collective practical knowledge? What individual practical knowledge is required for the group members?

I use the theory of 'seeing to it that', abbreviated to STIT, extended with epistemic notions, to study the concepts of collective and individual *know-how*

and their interrelation. Simply stated, if an individual agent knows how to φ then this is witnessed by a refinement ψ where, furthermore, she knows that she can φ by ψ-ing and she knows that she can knowingly ψ. For example, I know how to cook risotto by going through the steps of a recipe I found online. This highlights that my theory of individual practical knowledge relies on the concepts of *action hierarchies*, as when one is φ-ing by ψ-ing, and *knowingly doing*. Similarly, if a group of agents knows how to jointly φ, then this is witnessed by a refinement which specifies each member's part, where, furthermore, they collectively know that they can see to it that φ by each member playing her corresponding part and they collectively know that each member knows how to play her part.

We have seen in the previous chapter that in case communication is possible and agreement unproblematic, then responsibility voids can be prevented by adopting a good group plan. It is therefore important to highlight that I currently focus on cases where communication is impossible and agreement is problematic.[1] Coordination games illustrate an important way in which collective know-how may be undermined, since it may be impossible for a group to know how to coordinate (see, for example, Figure 1.1). With a focus on these coordination problems and action hierarchies, my central definitions clarify the complex concepts of individual and collective know-how. My central theorems concern the *characterization* of both individual and collective know-how. On the one hand, I show that an individual agent knows how to do something if and only if the agent knows that it is possible for her to knowingly do it. Collective know-how, on the other hand, requires common knowledge of an *effective division* of tasks among the group's members and of the fact that each member knows how to carry out her part. Accordingly, if a group aims to achieve some collective goal

[1] At the end of § 3.4 I briefly discuss some implications of my theory of collective know-how for cases where agreement is unproblematic and communication is possible.

and collectively knows how to do it, then no communication is needed: if each member knowingly plays her part (which she knows how to do), then the group jointly achieves their collective goal.

I illustrate my STIT formalism with ideas and concepts from the philosophical, artificial intelligence, and economics literature on action, knowledge, ability, practical knowledge, and joint action. This serves four purposes. First, the concepts and ideas from this multi-disciplinary literature help to clarify what my formalism is meant to model and yields a partial justification for my analysis of individual and collective know-how, and their interrelation. Second, the formal framework helps to clarify and specify elements of my conceptual analysis, most importantly, action hierarchies, knowingly doing, and effective divisions of tasks among the group's members. The concept of knowing-how is complex, so my formalization is aimed at reducing the ambiguities. Third, in interdependent scenarios it is crucial to reason about the collective capacities of you and your fellows. The formal framework offers an initial step to applications in artificial intelligent agents (§ 3.5). Fourth, my formalism captures important aspects of concepts and ideas developed in the philosophical, artificial intelligence, and economics literature and is therefore of central relevance for debates on joint action, practical knowledge, and collective intentionality. To highlight this, I briefly discuss the implications of my characterization of individual know-how for the debates on intellectualism (§ 3.3) and subjective obligations (§ 3.6), and I explore the relation between collective and individual blameworthiness in light of my theory of collective know-how (§ 3.6).

The chapter is vast, so it is of utmost importance to be guided by a discussion of its outline. In § 3.2, I present an epistemic extension of the basic STIT framework that will be the backbone for my theory of know-how and I provide a complete axiomatization for an epistemic STIT language. Those familiar with STIT theory and epistemic STIT theory may decide to skip this preliminary section. To pave

the way for building my theory of individual practical knowledge, I discuss action hierarchies in light of the philosophical literature on action individuation (§ 3.3.1) and present the concept of knowingly doing, which will be extensively contrasted with the artificial intelligence and economics literature on knowledge, action, and ability (§ 3.3.2). I then define implicit and explicit individual know-how and prove the central theorems on individual know-how: individual know-how can be characterized in my formalism, despite the fact that the distinction between so-called implicit and explicit know-how cannot be characterized (§ 3.3.3). A discussion of the intellectualism/anti-intellectualism debate follows. In § 3.4, I extensively discuss my intuitions regarding collective know-how, present a definition of it motivated by these discussions, and prove that it can be characterized. There follows a discussion highlighting a connection with the previous chapter: the concept of collective know-how includes an interchangeable group plan. In § 3.5, I discuss the relation of my work to artificial intelligence research. Most importantly, I argue that the characterization of practical knowledge – individual and collective – is important for the agent-based approach to artificial intelligence. I link and contrast my account with some dominant views in the literature on logics for knowledge and action, and, finally, I show how my views connect with ideas from the hierarchical planning literature. In the concluding section, I use my formalism to study aspects of subjective obligations and collective backward-looking responsibility.

3.2 Epistemic STIT Theory

To characterize my views concerning collective know-how in a simple way, I use an epistemic extension of the basic STIT framework. (See §§ 1.2–1.3 for a more detailed introduction and discussion of the basic, non-epistemic, STIT framework and its connections to game theory.) The framework I put forward here is a standard epistemic extension of the simple STIT models discussed in § 1.2 and,

more specifically, it is related to and inspired by the framework developed by Broersen (2011a). In this section I introduce this formal framework and present a completeness result. A discussion of the interpretation of the formal framework is postponed to the next section.

My epistemic STIT framework is best thought of as representing the possibilities, knowledge, and group actions at a single moment in time. One may think that an adequate theory of know-how requires a strategic context. That is, one may think that know-how needs to be situated in extensive form games or temporally extended action, rather than in one-shot games or instantaneous action. Although such a context may contribute to a theory of know-how, the additional complexity is unnecessary for the difficulties I presently raise and address. Moreover, it seems that my considerations are straightforwardly transferable to *strategic* STIT frameworks (see, for instance, Herzig and Troquard, 2006; Broersen et al., 2006; Broersen and Herzig, 2015; Duijf and Broersen, 2016).

The modal language of epistemic STIT includes four operators. The formulas in this language are evaluated at *dynamic states* (to guide the readers' intuitions they consist of a state and a history).[2] It is important to note that a dynamic state is taken to include the complete temporal evolution of the world. This means that the truth of, for instance, a temporal condition may depend on the exact history of evaluation. Our modal language, first, includes a historical necessity operator $\Box\varphi$, which expresses that φ holds at the current state, regardless of how the future unfolds. Second, it includes an agency operator $[\mathcal{H}\ \mathsf{stit}]\varphi$, which expresses that the group $\mathcal{H}$ sees to it that φ. Or, equivalently, that group $\mathcal{H}$ guarantees that φ holds, regardless of what the others do. Finally, the epistemic operators $\mathsf{K}_i\varphi$ and $\mathsf{C}_{\mathcal{H}}\varphi$ express that agent i knows that φ and that the group $\mathcal{H}$ commonly knows that φ.[3]

[2] These dynamic states correspond to moment/history pairs in traditional STIT models (§ 1.2).
[3] Xu (2015) presents an excellent survey on STIT theory and its epistemic extensions.

Definition 3.1 (Syntax)**.** *The formal language $\mathfrak{L}_{ESTIT}$ is:*

$$\varphi \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\mathcal{H} \text{ stit}]\varphi \mid \mathsf{K}_i\varphi \mid \mathsf{C}_{\mathcal{H}}\varphi,$$

where p ranges over a given countable set of propositions P, i ranges over a given finite set of agents Ags, and $\mathcal{H}$ ranges over subsets of Ags.

I use standard abbreviations for duals of modal operators, i.e. $\Diamond\varphi \equiv \neg\Box\neg\varphi$, $\langle\mathcal{H} \text{ stit}\rangle\varphi \equiv \neg[\mathcal{H} \text{ stit}]\neg\varphi$, and $\hat{\mathsf{K}}_i\varphi \equiv \neg\mathsf{K}_i\neg\varphi$. $\mathsf{C}_{\mathcal{H}}$ is the standard common knowledge operator (see Fagin et al., 2003; Meyer and van der Hoek, 1995), and I abbreviate $\bigwedge_{i\in\mathcal{H}} \mathsf{K}_i\varphi$ by $\mathsf{E}_{\mathcal{H}}\varphi$, which is the standard operator for 'everybody in $\mathcal{H}$ knows φ'. Common knowledge $\mathsf{C}_{\mathcal{H}}\varphi$ intuitively corresponds to everybody in $\mathcal{H}$ knows φ, everybody in $\mathcal{H}$ knows that everybody in $\mathcal{H}$ knows φ, and so on ad infinitum.

An epistemic STIT model is an extension of a branching-time agency model: it is supplemented with indistinguishability relations $\sim_i$, one for each agent $i \in Ags$.[4] The models that I use are mostly standard, which makes way for an axiomatization (Theorem 3.1). The only non-standard aspect is the two-dimensionality of dynamic states, which consist of a history and a state.

Definition 3.2 (Epistemic STIT Models)**.** *An* epistemic STIT frame *is a tuple $\langle S, H, (Act^s_{\mathcal{H}}), (\sim_i)\rangle$, involving a set of states S, a set of histories $H \subseteq 2^S$,[5] a given finite set of agents Ags, for each group of agents $\mathcal{H} \subseteq Ags$ and each state $s \in S$ a finite set of actions $Act^s_{\mathcal{H}} \subset 2^{H_s}$ available to group $\mathcal{H}$ at s, and for each agent $i \in Ags$ an indistinguishability relation $\sim_i$, satisfying the following:*

- *for every state $s \in S$ it holds that $\langle W_s, Ags, (Act^s_{\mathcal{H}})\rangle$ is a STIT model, where $W_s = \{\langle s, h\rangle \in S \times H \mid s \in h\}$ (see Definition 1.10);*

[4] In artificial intelligence research such indistinguishability relations are standardly used to represent an agent's knowledge (see Meyer and van der Hoek, 1995; Fagin et al., 2003). They straightforwardly correspond to partition structures that are commonly employed in game theory and economics to model the information states of the players (see Aumann, 1999).

[5] Histories can be viewed as infinite paths in Alternating-time Temporal Logic, standardly abbreviated to ATL, frameworks (cf. Alur et al., 2002).

- *for every $i \in Ags$ the indistinguishability relation $\sim_i$ is an equivalence relation on the set of dynamic states $\bigcup_s W_s$.*

For each group $\mathcal{H}$, we let $\sim^*_{\mathcal{H}}$ denote the reflexive transitive closure of the union of the relations $\{\sim_i | \ i \in \mathcal{H}\}$, which will be used to interpret the common knowledge operator.

Such an epistemic STIT frame is extended to an epistemic STIT model *by supplementing a valuation $V : \mathcal{P} \to 2^{S \times H}$, which assigns to each atomic proposition p the set of dynamic states $V(p)$ in which it is true.*

It is essential to note that this notion of knowledge differs crucially from existing accounts in the literature on knowledge and action. In the ATL tradition, imperfect information is usually modelled using an epistemic indistinguishability relation on *static* states, rather than on dynamic states (see van der Hoek and Wooldridge, 2003). In such a model, knowledge only concerns static conditions, not actions themselves. In contrast, because I let epistemic indistinguishability concern dynamic states, the formalism is able to express knowledge of what agents are doing. In the next section I will extensively discuss the relation of my framework to the artificial intelligence and economics literature (§ 3.3.2).

The truth conditions for the semantics of the logical operators are standard. The non-standard aspect is the two-dimensionality of the semantics, that is, the truth value is given for *dynamic* states consisting of a history and a static state.

Definition 3.3 (Semantics)**.** *Let $\mathcal{M} = \langle S, H, (Act^s_{\mathcal{H}}), (\sim_i), V \rangle$ be an epistemic STIT model. Then the truth of a formula φ at a dynamic state $\langle s, h \rangle$ in $\mathcal{M}$, notation: $\mathcal{M}, \langle s, h \rangle \vDash \varphi$, is given by (suppressing the standard propositional clauses):*

$\mathcal{M}, \langle s, h \rangle \vDash \Box\varphi$ *iff* *for every dynamic state $\langle s', h' \rangle$ satisfying $s = s'$ it holds that $\mathcal{M}, \langle s', h' \rangle \vDash \varphi$;*

$\mathcal{M}, \langle s, h \rangle \vDash [\mathcal{H}\ \mathsf{stit}]\varphi$ *iff* *for every dynamic state $\langle s', h' \rangle$ satisfying $s = s'$ and $h' \in Act^s_{\mathcal{H}}(h)$ it holds that $\mathcal{M}, \langle s', h' \rangle \vDash \varphi$;*

$\mathcal{M}, \langle s, h\rangle \vDash \mathsf{K}_i\varphi$ *iff* *for every dynamic state* $\langle s', h'\rangle$ *satisfying* $\langle s, h\rangle \sim_i \langle s', h'\rangle$ *it holds that* $\mathcal{M}, \langle s', h'\rangle \vDash \varphi$;

$\mathcal{M}, \langle s, h\rangle \vDash \mathsf{C}_{\mathcal{H}}\varphi$ *iff* *for every dynamic state* $\langle s', h'\rangle$ *satisfying* $\langle s, h\rangle \sim^*_{\mathcal{H}} \langle s', h'\rangle$ *it holds that* $\mathcal{M}, \langle s', h'\rangle \vDash \varphi$.

Validity on a model, notation: $\mathcal{M} \vDash \varphi$*, validity on a frame, notation:* $\mathcal{F} \vDash \varphi$*, and general validity, notation:* $\vDash \varphi$*, are defined as usual.*

Given a model $\mathcal{M}$ *and a static state* $s \in S$*, the truth set of a formula* φ *relative to* $\mathcal{M}$ *or to* s*, notation:* $[\![\varphi]\!]_{\mathcal{M}}$ *and* $[\![\varphi]\!]_s$ *respectively, is given by* $\{\langle s, h\rangle \in \mathcal{M} \mid \mathcal{M}, \langle s, h\rangle \vDash \varphi\}$*, where either the model* $\mathcal{M}$ *is fixed, or the static state* s *and the model* $\mathcal{M}$ *are fixed.*

Definition 3.2 says that, like in standard STIT semantics, dynamic states based on the same static state can have different valuations of atomic propositions. This captures the Ockhamistic view that the present static state does not determine the truth of every proposition. The intuitive reason for positing alternative histories through a particular static state is that they witness different actions that the agents can perform. On this view, two histories branch at the current static state if the agents currently perform different actions along these histories.[6] These histories constitute different dynamic states in combination with the current static state. So, in particular, agentive conditions, expressed by $[i\ \mathsf{stit}]\varphi$, may evaluate to different truth values at these dynamic states.

A complete logic is readily available for these epistemic STIT models, because they yield a fusion of STIT logic and standard epistemic logic.[7] This complete logic is important for two main reasons: (1) it helps our conceptual analysis, and clarifies

[6] Note the absence of the 'only if' here, which only holds for deterministic models (see Definition 1.12).

[7] Herzig and Schwarzentruber (2008) prove that group STIT is non-axiomatizable, so this seems at odds with my axiomatization result. The key to my axiomatization is that, in contrast to Herzig and Schwarzentruber, I do not impose the intersection property (see Definition 1.12 in § 1.3).

the relations between the introduced modalities; (2) it enables artificial intelligent agents to reason about abilities, knowledge, and know-how, both individual and collective, using the logical system. (All claims are proved in Appendix C.)[8]

Theorem 3.1 (Completeness Epistemic STIT)**.** *The following axiom schemas, in combination with a standard axiomatization for propositional logic, and the standard rules* (*such as necessitation*) *for the normal modal operators, provide a complete Hilbert system for the validities on epistemic STIT models:*

(S5 Historical Necessity)	*S5 for* $\Box$
(S5 Agency)	*for each group* $\mathcal{H}$*: S5 for* $[\mathcal{H}\ \mathsf{stit}]$
(Agent Monotonicity)	*for all groups* $\mathcal{F}$ *and* $\mathcal{G}$ *satisfying* $\mathcal{F} \subseteq \mathcal{G}$*:* $[\mathcal{F}\ \mathsf{stit}]\varphi \rightarrow [\mathcal{G}\ \mathsf{stit}]\varphi$
(Independence of Agency)	*for all groups* $\mathcal{F}$ *and* $\mathcal{G}$ *satisfying* $\mathcal{F} \cap \mathcal{G} = \emptyset$*:* $\Diamond[\mathcal{F}\ \mathsf{stit}]\varphi \wedge \Diamond[\mathcal{G}\ \mathsf{stit}]\psi \rightarrow \Diamond([\mathcal{F}\ \mathsf{stit}]\varphi \wedge [\mathcal{G}\ \mathsf{stit}]\psi)$
(S5 Knowledge)	*for each* $i \in Ags$*: S5 for* K_i
(Public Knowledge)	*for each group* $\mathcal{H}$*:* $\mathsf{C}_{\mathcal{H}}\varphi \rightarrow (\varphi \wedge \mathsf{E}_{\mathcal{H}}\mathsf{C}_{\mathcal{H}}\varphi)$
(Induction)	*for each group* $\mathcal{H}$*:* $\varphi \wedge \mathsf{C}_{\mathcal{H}}(\varphi \rightarrow \mathsf{E}_{\mathcal{H}}\varphi) \rightarrow \mathsf{C}_{\mathcal{H}}\varphi$

3.3 Individual Practical Knowledge

I start in § 3.3.1 by explicating my views on action hierarchies in light of the philosophical literature on action individuation. In § 3.3.2 I develop an account of

[8]Some comments on the logical system. For those unfamiliar with modal logic, it may be useful to point out that one of the basic results is that the S5 axioms (reflexivity: $\Box\varphi \rightarrow \varphi$, symmetry: $\varphi \rightarrow \Box\Diamond\varphi$, and transitivity: $\Box\varphi \rightarrow \Box\Box\varphi$) jointly correspond to an accessibility relation that is an equivalence relation. Moreover, an equivalence relation corresponds to a partitioning. I refer the reader to the standard textbook treatment of modal logic (Blackburn et al., 2001). The standard interpretation for the knowledge operator is as follows: (i) reflexivity, $\mathsf{K}_i\varphi \rightarrow \varphi$, requires that knowledge is factive, (ii) transitivity, $\mathsf{K}_i\varphi \rightarrow \mathsf{K}_i\mathsf{K}_i\varphi$, requires that knowledge is positively introspective, and (iii) euclidity, $\neg\mathsf{K}_i\varphi \rightarrow \mathsf{K}_i\neg\mathsf{K}_i\varphi$, requires that knowledge is negatively introspective.

This axiomatization of common knowledge is from the work of Meyer and van der Hoek (1995). Other axiomatizations of common knowledge rely on an induction rule such as $\varphi \rightarrow \mathsf{E}_{\mathcal{H}}(\varphi \wedge \psi) \mid \varphi \rightarrow \mathsf{C}_{\mathcal{H}}\psi$ (Fagin et al., 2003) or $\varphi \rightarrow \mathsf{E}_{\mathcal{H}}\varphi \mid \mathsf{E}_{\mathcal{H}}\varphi \rightarrow \mathsf{C}_{\mathcal{H}}\varphi$ (Lismont, 1993).

knowingly doing with an eye on the literature on formalisms of knowledge and action. This will enable a comparison of various modes of knowledge to the ex ante/ex interim/ex post distinction common in game theory, and a discussion of action types and uniform strategies. Finally, in § 3.3.3 I discuss individual know-how, provide some (im)possibility results, and eventually give a characterization of individual know-how. I finish with a discussion of the intellectualism/anti-intellectualism debate from the perspective of my theory of individual know-how.

3.3.1 Action Hierarchies

Much ink has been spilled on the concepts of knowing-how, actions, and ability. To explicate my ideas and to illustrate my interpretation of STIT theory, I will explain the concept of *action hierarchies* with reference to the relevant philosophical literature.[9] Elizabeth Anscombe, in her seminal work, writes:

> Are we to say that the man who (intentionally) moves his arm, operates the pump, replenishes the water-supply, poisons the inhabitants, is performing *four* actions? Or only one? (Anscombe, 1963, p. 45)[10]

Are there as many actions (and as many intentions) as there are such descriptions? I follow Anscombe (1963) and Davidson (1963) and submit that there is but one action which is described in various ways (albeit perhaps not by a definite description).[11] In a STIT model, it is obviously true that at each dynamic state $\langle s, h\rangle$ an agent is only performing one particular action, viz. the action represented by

[9] In § 3.5 I briefly elaborate on connections between my view on action hierarchies and the view in the artificial intelligence literature on planning. More specifically, I point out that my framework naturally relates to hierarchical planning (§ 3.5.1).

[10] Anscombe (1963, p. 46) answers her question as follows: "In short, the only distinct action of his that is in question is this one, A. For moving his arm up and down with his finger round the pump handle is, in these circumstances, operating the pump; and, in these circumstances, it is replenishing the house water-supply; and, in these circumstances, it is poisoning the household."

[11] This way of answering the question makes me a so-called 'coarse-grained theorist'. Goldman (1971) is a classic example of the opposite 'fine-grained theorist', who, among other things, maintains that the man in the example performs more than one action.

$Act_i^s(h)$. It is best to interpret the expression $[i\ \mathsf{stit}]\varphi$ as describing agent i's action as one that guarantees φ.[12] Having clarified this distinction, I should point out that I occasionally conflate $[i\ \mathsf{stit}]\varphi$ as an act description and as a φ-action.

Often, the expression $[i\ \mathsf{stit}]\varphi$ applies to multiple acts rather than to a definite unique action. This means that various descriptions, such as $[i\ \mathsf{stit}]\varphi$ and $[i\ \mathsf{stit}]\psi$, may pick out different *sets* of acts. In the example from Anscombe, for instance, it seems natural to say that the man is pumping the water by moving his arm, yet the converse does not hold. How should we interpret the by-relation between these particular two act descriptions?[13] In such a case I submit that the man's arm-movement is a *refinement* of his water-pumping. The example thus gives rise to an action hierarchy which contains several refinements (or granularities) of act descriptions.[14]

What is the by-relation between act descriptions? There are various guises that are standardly distinguished in philosophy, e.g. causal, intentional, teleological, constitutional, etc.[15] My formalism naturally grants a conception of the by-relation that relies on historical necessity: if, relative to a dynamic state $\langle s, h\rangle$, it is historically necessary that agent i sees to it that ψ implies that agent i sees to it that φ, then we say that agent i can φ by ψ-ing.[16] Semantically, this means

[12] This entails that someone who considers the semantical framework as primary will be a coarse-grained theorist, while someone who takes the syntactical framework as primary may be open to the fine-grained theory of action individuation. A semantical fine-grained theory may be formalized by letting $Act_i^s(h)$ be a collection of subsets of H_s rather than a subset of H_s. In fact, in STIT theory Broersen (2009) pioneered this way of modelling strategic action: a strategy is equated with the set of histories compatible with it; and a certain history may be compatible with multiple strategies.

[13] Anscombe (1963, see p. 45) associates these descriptions with the man's intentions. I will not do so; instead my proposal is best viewed in terms of agency-causation. That is, $[i\ \mathsf{stit}]\varphi$ means that agent i causes φ rather than agent i intentionally φs. Intentions are studied in Chapter 4.

[14] Feinberg (1970, pp. 119–151) has referred to this, or a similar, idea as the "accordion effect". Davidson (1980, p. 53) explores the idea that "we may take the accordion effect as a mark of agency".

[15] On my view, this issue is connected to the literature on grounding (see the SEP entry by Bliss and Trogdon, 2016). Epstein (2015) gives a uniform framework for modelling all these senses of grounding. He convincingly shows that all these senses may be modelled as a necessary implication, although they may vary in the types of facts that ground this relation, for instance, some are grounded in frame conditions, others in model properties.

[16] Compare Wilson (1989), who presents a teleological account of intentionality, where 'agent i ψ'd in order to φ' is analysed as 'agent i ψ'd because he wanted to φ and believed that $\mathrm{By}(\psi, \varphi)$'.

that the actions in Act_i^s that guarantee ψ also guarantee φ. Or, equivalently, $[\![[i \text{ stit}]\psi]\!]_s \subseteq [\![[i \text{ stit}]\varphi]\!]_s$.[17] We say that this is a refinement because every way for agent i to see to it that ψ is a way for her to see to it that φ.

Definition 3.4 (Refinements & Action Hierarchies)**.** *Let $\mathcal{M} = \langle S, H, (Act^s_{\mathcal{H}}), (\sim_i), V\rangle$ be an epistemic STIT model, let $\mathcal{H} \subseteq Ags$ be a group, and let $\langle s, h\rangle$ be a dynamic state. Then we say that, relative to $\langle s, h\rangle$, ψ* is a refinement of φ for $\mathcal{H}$ *if and only if it is historically necessary that if $\mathcal{H}$ sees to it that ψ then $\mathcal{H}$ sees to it that φ, that is,*

$$\mathcal{M}, \langle s, h\rangle \vDash \Box([\mathcal{H} \text{ stit}]\psi \to [\mathcal{H} \text{ stit}]\varphi).$$

When the converse does not hold, we call it a proper refinement.[18]

It is useful to add that whenever an individual agent i brings about φ by bringing about ψ, then bringing about ψ is a way for her to bring about φ. There may be various ways in which she can bring about φ, so the converse need not hold. This interpretation is naturally included in my informal views on know-how: an agent knows how to φ if and only if there is a refinement ψ which she knows is a way for her to φ.

3.3.2 Knowingly Doing

My theory of know-how is cast against the background of a theory of *knowingly doing*. Jan Broersen (2011a, § 3) gives a first logical treatment of knowingly doing, which I largely follow. With the epistemic STIT framework at play (§ 3.2), we can express that agent *i knowingly* sees to it that φ by using the formula $\mathsf{K}_i[i \text{ stit}]\varphi$. The semantics of this expression are given in terms of epistemic indistinguishability

[17] Note the important difference with $[\![\psi]\!]_s \subseteq [\![\varphi]\!]_s$ and $[\![[i \text{ stit}]\psi]\!]_{\mathcal{M}} \subseteq [\![[i \text{ stit}]\varphi]\!]_{\mathcal{M}}$. The former drops the $[i \text{ stit}]$-operator and therefore misses the fact that the by-relation applies to act descriptions. The latter expresses a global by-relation rather than a local by-relation, that is, it says that *whenever* i sees to it that ψ then she sees to it that φ rather than *at the current moment* if i sees to it that ψ then she sees to it that φ.

[18] Note that the syntactical counterparts of $[\![\psi]\!]_s \subseteq [\![\varphi]\!]_s$ and $[\![[i \text{ stit}]\psi]\!]_{\mathcal{M}} \subseteq [\![[i \text{ stit}]\varphi]\!]_{\mathcal{M}}$, mentioned in the previous footnote, are $\mathcal{M}, \langle s, h\rangle \vDash \Box(\psi \to \varphi)$ and $\mathcal{M} \vDash \Box([\mathcal{H} \text{ stit}]\psi \to [\mathcal{H} \text{ stit}]\varphi)$.

relations on *dynamic* states. An agent i knowingly does φ if the formula $[i\ \mathsf{stit}]\varphi$ holds for all the dynamic states in the epistemic equivalence set containing the actual dynamic state. In light of the previous discussion on action hierarchies, this means that agent i knows that her current action can be described as one guaranteeing φ. That is, an agent knowingly does φ if and only if the agent knows that she guarantees that φ holds, regardless of what the other agents do. Or, equivalently, φ is something that an agent knowingly does if and only if she is certain that the action she performs will result in φ.

To illustrate that knowingly doing φ is different from simply seeing to it that φ holds, let us consider a simple example. Imagine a deck of cards, spread out face down on a table in front of Ann, and suppose she can only pick one card. Furthermore, imagine that Zach simultaneously writes down the name of a card, for instance, jack of hearts, on a piece of paper. Suppose that Zach writes down 'jack of hearts' and that Ann randomly picks the jack of hearts. Let φ_m express that the card Ann picks *matches* what Zach writes down, and let a and z stand for Ann and Zach, respectively.[19] A simplification of the scenario is depicted in Figure 3.1. Ann does not know which card she picks because she cannot distinguish, for instance, a jack of hearts from an ace of hearts. This is represented by letting $K_a^{\heartsuit A}$ and $K_a^{\heartsuit J}$ be part of the same information partition. However, Ann is, as a matter of fact, guaranteeing that she picks a jack of hearts. Formally, this is expressed by $[a\ \mathsf{stit}]\varphi_{\heartsuit J}$. She does so unknowingly, as is expressed by $\neg \mathsf{K}_a[a\ \mathsf{stit}]\varphi_{\heartsuit J}$. Finally, note that the card Ann picks matches what Zach writes down even though Ann does not knowingly do it and she also does not guarantee it. That is, φ_m holds even though $\mathsf{K}_a[a\ \mathsf{stit}]\varphi_m$ and $[a\ \mathsf{stit}]\varphi_m$ do not hold.

To highlight the flexibility and interpretation of my framework I will now discuss two possible properties of knowingly doing. These properties are in the

[19] Moreover, from here on, formulas φ and ψ are subscripted with, for instance, $\spadesuit$ and J to make clear that I am referring to picking a spade and a jack. So $\varphi_{\spadesuit A}$ and $\varphi_{\heartsuit J}$ concern picking an ace of spades and a jack of hearts, respectively.

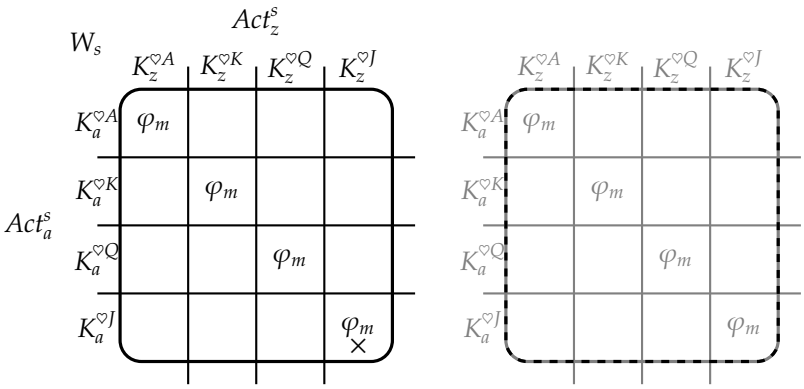


Figure 3.1: *A simplified epistemic STIT model that represents Ann's decision problem: Act_a^s and Act_z^s represent the actions available to Ann and Zach respectively; × marks the actual dynamic state; and Ann's epistemic state is depicted on the right-hand side, where the epistemic indistinguishability relation is given by the universal relation on W_s.*

Sahlqvist class, which means that they yield a complete logic. I will present them as axioms in the epistemic STIT language (Definition 3.1) and give the corresponding first-order conditions on epistemic STIT frames (Definition 3.2).[20]

Proposition 3.1.

(OAC) *The* own-action condition*, schematically expressed by* $\mathsf{K}_i\varphi \rightarrow \mathsf{K}_i[i\ \mathsf{stit}]\varphi$*, corresponds to the condition:*

For all dynamic states $\langle s,h\rangle$*,* $\langle s',h_1\rangle$*, and* $\langle s',h_2\rangle$*, if* $\langle s,h\rangle \sim_i \langle s',h_1\rangle$ *and* $h_2 \in Act_i^{s'}(h_1)$ *then* $\langle s,h\rangle \sim_i \langle s',h_2\rangle$*.*

(Unif-H) *The* uniformity of historical possibility property*, schematically expressed by* $\Diamond\mathsf{K}_i\varphi \rightarrow \mathsf{K}_i\Diamond\varphi$*, corresponds to the confluency condition:*

[20] These, and further, correspondences can be algorithmically checked using the SQEMA algorithm (Conradie et al., 2006).

Broersen (2011a, Propositions 3.1 and 3.2) uses a STIT logic about affecting *next* states, so-called XSTIT, to model knowingly doing and discusses two additional properties: (i) knowledge about next states, which expresses that "agents cannot know more about next states than what is affected by the choices they have" (p. 145) and (ii) effect recollection, which expresses that "the effects of an action that is knowingly performed are known in the next state" (p. 145 – it is a dynamic version of perfect recall).

> *For all dynamic states* $\langle s, h_1\rangle$, $\langle s, h_2\rangle$, *and* $\langle s', h'_1\rangle$, *if* $\langle s, h_1\rangle \sim_i \langle s', h'_1\rangle$ *then there is a history* h'_2 *such that* $\langle s, h_2\rangle \sim_i \langle s', h'_2\rangle$.

Before discussing the interpretation of these conditions and their appeal, I would like to briefly point out whether my theories of individual and collective know-how (§§ 3.3.3–3.4) depend on these conditions. The own-action condition plays no role in my theorizing; in fact I even think that it is undesirable (as will be hinted at below). At the moment, I am unsure whether the uniformity of historical possibility property (Unif-H) is a desirable or natural property to impose on the models. But, because game theorists seem to take it for granted, I will follow suit.[21] I will clearly indicate when (Unif-H) plays a role in proving a result.

It may be useful to add that, given that the necessitation-rule and both the transitivity- and the K-axiom hold for K_i, the own-action condition is equivalent to $\langle i \ \mathsf{stit}\rangle\varphi \rightarrow \hat{\mathsf{K}}_i\varphi$. The own-action condition therefore corresponds to the simple condition that for any dynamic states $\langle s, h\rangle$, $\langle s, h'\rangle$ if $h' \in Act_i^s(h)$ then $\langle s, h\rangle \sim_i \langle s, h'\rangle$. That is, any information state in the partition induced by $\sim_i$ is a union of actions from Act_i.

The own-action condition expresses the idea that an agent cannot know *more* about the current dynamic state than what she knows about what she herself brings about. Or, equivalently, when an agent does not know that she sees to it that φ then she also does not know that φ holds. In other words, agents can only know something about the current dynamic state if it is the result of an action they themselves knowingly perform. Conversely, an agent *unknowingly* does φ if and only if (i) she is performing an action that guarantees φ and (ii) she considers it possible that φ does not obtain. Jan Broersen writes:

> In general the things an agent does unknowingly vastly outnumber the things an agent *knows* it does. For instance, by sending an email,

[21] An elaborate discussion of my worries has to be postponed to another occasion.

> we may enforce many, many things we are not aware of, which are nevertheless the result of sending the email. All these things we do *unknowingly* by knowingly sending the email. (Broersen, 2011a, p. 145)

We might say that the own-action condition is a consequence of the assumption that agents cannot know what actions other agents perform concurrently.[22] The independence-of-agency condition (see Definitions 1.8 and 1.10, and Theorem 3.1) guarantees that choices of agents always refine choices of other agents. So if an agent knows about the choices of other agents, then she knows more about the future than what is guaranteed by her own choice.

The uniformity of historical possibility property expresses the idea that whenever it is possible for agent i to know φ then she knows that φ is possible. It is easy to see that the confluency condition also corresponds to $\mathsf{K}_i\Box\varphi \rightarrow \Box\mathsf{K}_i\varphi$, or, equivalently, $\Diamond\hat{\mathsf{K}}_i\varphi \rightarrow \hat{\mathsf{K}}_i\Diamond\varphi$. This means that whenever an agent knows that φ is historically settled, then it is historically settled that she knows φ. When we think of histories as witnessing different actions that the agents can perform, then this means that an agent knows that φ holds regardless of the actions everyone performs only if she knows φ regardless of the actions everyone performs. Or, for the initial formulation in Proposition 3.1, whenever the agents can perform certain actions that are compatible with the fact that agent i knows φ, then agent i knows that the agents can perform certain actions that are compatible with φ.

3.3.2.1 Ex Ante, Interim, and Ex Post Knowledge

It is useful to discuss the notions of knowledge and knowingly doing in my formalism in light of a distinction made by economists:

[22] It is questionable whether this property also holds under the assumption of common knowledge of rationality, which is standardly assumed in game theory. Under this assumption, it seems appropriate to say that an agent knows that her opponents are avoiding a ridiculous, yet possible, action. This would mean that she considers it impossible that her current action is compatible with such ridiculous actions of her opponents yet it is (historically) possible that her opponents perform these actions.

> Economists distinguish three stages in differential information environments. Ex ante, no one has any private information; in the interim, each agent has his private information only; ex post, all information is revealed to all. In our context, ex ante the protagonist knows only the belief system B – which is commonly known by all players; in the interim, each player knows his type, but not those of the others; ex post, each player knows the types of all players. (Aumann and Dreze, 2008, p. 80)[23]

It thus seems natural to suppose that ex post knowledge refines interim knowledge, which in turn refines ex ante knowledge.[24] That is, whenever an agent knows φ ex ante then she knows φ ex interim, and, similarly, when an agent knows ψ ex interim then she knows ψ ex post.[25]

Since knowingly doing is characterized using the knowledge operator, the view on knowledge captured in my formalism is best interpreted as interim knowledge. That is, it is best viewed as interim knowledge because knowingly doing, expressed by $\mathsf{K}_i[i\ \mathsf{stit}]\varphi$, is a concept that clearly operates at the interim stage of decision-making. As mentioned before, whenever $\mathsf{K}_i[i\ \mathsf{stit}]\varphi$ holds at a dynamic state $\langle s, h\rangle$ then it is the case that agent i knows that she ensures that φ holds, regardless of what the other agents do. Accordingly, $\mathsf{K}_i\varphi$ might be interpreted as expressing that agent i knows that φ holds, regardless of what *she thinks* the other agents might do. When the own-action condition is violated,

[23] Lorini et al. (2014, p. 1314) write: "Each type of knowledge is defined with respect to the time of the agent's choice: before one's choice (*ex ante* knowledge), after one's choice but before knowing the choices of others (*interim* knowledge), and after the choices of all agents have been made public (*ex post* knowledge)." Compare Horty and Pacuit (2017, pp. 31–32): "an agent's ex ante knowledge is the information available to the agent without taking into account any actions she is currently executing, while the agent's ex interim knowledge is information that does take into account whatever actions the agent is currently executing, along with the effects of these actions".

[24] Lorini et al. (2014, § 2.2.2) model agent i's ex ante, interim, and ex post knowledge using accessibility relations $\mathcal{E}_i^{\bullet\circ\circ}$, $\mathcal{E}_i^{\circ\bullet\circ}$, and $\mathcal{E}_i^{\circ\circ\bullet}$, respectively. The idea that these are refinements is expressed by the inclusions $\mathcal{E}_i^{\bullet\circ\circ} \supseteq \mathcal{E}_i^{\circ\bullet\circ} \supseteq \mathcal{E}_i^{\circ\circ\bullet}$.

[25] The assumption that ex post knowledge refines ex ante knowledge is standardly referred to as 'perfect recall'.

knowingly doing φ may require that φ holds even at epistemically inaccessible dynamic states. It thus seems appropriate to submit that $\mathsf{K}_i\varphi$ expresses that agent i interim knows φ.

The own-action condition should not be confused with a standard assumption in epistemic game theory: knowledge of one's own action . According to Pacuit and Roy (2017), knowledge of one's own action is "the trademark of ex-interim situations". It corresponds to the condition that for any dynamic states $\langle s, h\rangle$ and $\langle s, h'\rangle$, if $\langle s, h\rangle \sim_i \langle s, h'\rangle$ then $h' \in Act_i^s(h)$. This means that an agent cannot be uncertain about which action she performs. In other words, an agent cannot know *less* than what she herself brings about.[26]

This reveals a conflicting assumption in game theory and my work on knowingly doing. I drop the standard assumption in game theory that an agent knows how to perform any of her available actions. To illustrate this, reconsider the face-down deck of cards example. On my account, it is impossible for Ann to knowingly pick the jack of hearts. It is, however, possible that she performs that action, despite unknowingly doing so. This conflicts with the standard assumption in epistemic game theory. According to game theorists either (i) Ann cannot perform the action of picking the jack of hearts, or (ii) it is possible for Ann to knowingly do so. It is important to bear this conflict in mind.[27]

Agent i's knowledge that is independent of the actual action she performs is expressed by $\Box\mathsf{K}_i\psi$. It is, nonetheless, unclear whether this corresponds to ex ante knowledge.[28] The following two considerations may clarify why. First, because ex ante knowledge is not private knowledge it seems natural to equate it with

[26] Aumann (1987, p. 8) writes: "Of course, a player always knows which decision he himself takes."

[27] It seems that Aumann had games of *perfect information* in mind, a term that became standard only after, and perhaps because of, his seminal works were published. A contemporary game theorist would reply to my worries expressed in the face-down deck of cards example by saying that the example is best modelled as a game of imperfect information. In games of imperfect information, an epistemic sense of ability is typically modelled using action types, rather than actions. We will discuss connections to action types shortly.

[28] Compare Horty and Pacuit (2017, p. 34 – notation adapted): "if the agent has ex interim knowledge that φ no matter which of her available actions she happens to execute, this entails that she must have ex ante knowledge that φ".

common knowledge. Consequently, to express ex ante knowledge it seems necessary to use the common knowledge operator C_{Ags}. Second, ex ante knowledge is the available public information *before* making a choice. Although $\Box\mathsf{C}_{Ags}\varphi$ would capture that the knowledge is independent of the current actions, it is unclear whether this is equivalent to saying that it represents the information *before* any choices have been made. Perhaps some particular information is independent of the choice made yet only obtains after a choice is made (for instance, the information that a choice has been made).

Ex post knowledge is the knowledge that is attained after everyone acted. Since my simple epistemic STIT models and language abstracted away from the temporal structure of the possible dynamic worlds, it seems that this cannot be modelled appropriately. One could use more elaborate branching-time agency models and syntax, including a next operator X, to represent ex post knowledge by $\mathsf{XK}_i\varphi$. Alternatively, the sentence "ex post, all information is revealed to all" in the above quote suggests that one might model ex post knowledge as distributed knowledge: $\mathsf{D}_{Ags}\varphi$.[29]

These three types of knowledge help to clarify the interpretation of the kind of knowledge used in my formalism. The previous discussion is not meant to settle the correct interpretation of these three types of knowledge.[30] Rather, it is supposed to utilize my formalism to map a space of possible interpretations and ambiguities.

[29] The distributed knowledge of a group, say $\mathcal{H}$, is typically modelled by a derived indistinguishability relation $\sim^{\mathsf{D}}_{\mathcal{H}} := \bigcap_{i\in\mathcal{H}} \sim_i$. The evaluation rule is as follows: $\mathcal{M}, \langle s,h\rangle \vDash \mathsf{D}_{\mathcal{H}}\varphi$ if and only if for all $\langle s',h'\rangle$ satisfying $\langle s,h\rangle \sim^{\mathsf{D}}_{\mathcal{H}} \langle s',h'\rangle$ it holds that $\mathcal{M}, \langle s',h'\rangle \vDash \varphi$. (Which is equivalent to requiring that for all $\langle s',h'\rangle$ satisfying $\langle s,h\rangle \sim_i \langle s',h'\rangle$, for every $i \in \mathcal{H}$, it holds that $\mathcal{M}, \langle s',h'\rangle \vDash \varphi$.)

[30] This is unproblematic since the game-theoretical interpretation is also not settled: "At one extreme is the ex ante stage where no decision has been made yet. The other extreme is the ex post stage where the choices of all players are openly disclosed. In between these two extremes is the ex interim stage where the players have made their decisions, but they are still uninformed about the decisions and intentions of the other players. These distinctions are not intended to be sharp. Rather, they describe various stages of information disclosure during the decision-making process" (Pacuit and Roy, 2017).

3.3.2.2 Uniform Strategies and Action Types

It is standard to distinguish between games of complete information and games of incomplete information. In the former, the structure of the game is commonly known, while in the latter this condition is relaxed. In the ATL tradition, imperfect information is usually modelled using an epistemic indistinguishability relation on static states, rather than dynamic states (see van der Hoek and Wooldridge, 2003). That is, it is assumed that all uncertainty is due to uncertainty regarding the static state.[31] To study the concept of epistemic ability in ATL, one typically relies on so-called *action types*. Simply stated, one submits that agent i is able, in an epistemic sense, to do φ if and only if there is an action type available to her that guarantees φ.

The concept of knowingly doing naturally relates to action types and uniform strategies. Uniform strategies and action types have a long tradition in the literature on knowledge and action (see van der Hoek and Wooldridge, 2003; Jamroga and van der Hoek, 2004; Herzig and Troquard, 2006; Schobbens, 2004; Jamroga and Ågotnes, 2007). Following Herzig and Troquard (2006), in epistemic STIT models, I could define the epistemic indistinguishability relations on *dynamic states* as follows: $\langle s_1, h_1\rangle \sim_i \langle s_2, h_2\rangle$ if and only if agent i performs the same action type at these dynamic states. That is, if and only if $Act_i^{s_1}(h_1)$ and $Act_i^{s_2}(h_2)$ are instances of the same action type.[32] In this case, $\mathsf{K}_i[i\ \mathsf{stit}]\varphi$ would hold at $\langle s, h\rangle$ if and only if the action type agent i performs at $\langle s, h\rangle$ ensures that φ obtains in every indistinguish-

[31] In comparison, Bradley and Drechsler (2014, p. 1225) study three types of uncertainty: ethical, option, and state uncertainty: "Ethical uncertainty arises if the agent cannot assign precise utilities to consequences. Option uncertainty arises when the agent does not know what precise consequence an act has at every state. Finally, state space uncertainty exists when the agent is unsure how to construct an exhaustive state space."

[32] Recently, Horty and Pacuit (2017) argued that STIT models *need to be extended with action types* to address certain puzzles about knowledge and action. I believe their argument is misguided due to their persistence that the indistinguishability relations are on *static* states. Engaging in this debate and working out the details would lead me to far astray.

able dynamic state, regardless of what the others do.[33] Accordingly, performing a uniform strategy that ensures that φ holds at every indistinguishable state in ATL models would correspond to knowingly doing φ in the corresponding epistemic STIT model.[34]

The uniformity of historical possibility property (Unif-H) stated in Proposition 3.1 then relates to the concept of a *uniform strategy*. If we substitute $[i\ \mathsf{stit}]\varphi$ for φ, the (Unif-H) formula turns into $\Diamond \mathsf{K}_i[i\ \mathsf{stit}]\varphi \rightarrow \mathsf{K}_i\Diamond[i\ \mathsf{stit}]\varphi$.[35] This says that if it is possible for an agent to knowingly see to it that φ, then she knows there is an action available to her that ensures φ. Under the previously mentioned correspondence between knowingly doing and uniform strategies, the intuition that uniformity means that the same action types should be available at indistinguishable states follows from (Unif-H) and S5-axioms for K_i: $\Diamond \mathsf{K}_i[i\ \mathsf{stit}]\varphi \rightarrow \mathsf{K}_i\Diamond \mathsf{K}_i[i\ \mathsf{stit}]\varphi$.[36] This formula expresses the idea that when it is possible for an agent to knowingly ensure φ then she knows that there is an action available to her that she knows ensures φ. Or, equivalently, if she does not know that there is a way for her to knowingly ensure φ, then it is impossible for her to knowingly ensure φ.

3.3.3 Individual Know-how

The debates concerning practical knowledge are ambiguous regarding the meaning of ability. So let me clarify what *I* mean by ability: the notion of ability adopted

[33] It took a while for the literature on knowledge and action to successfully model this. Jamroga and Ågotnes (2007) were the first to do this using constructive knowledge. See also our brief discussion in Duijf and Broersen (2016, p. 24).

[34] Broersen et al. (2006) show that ATL can be embedded into STIT theory. The central coalitional strategic ability operator of ATL, expressed by $\langle\langle \mathcal{H} \rangle\rangle \varphi$, corresponds to the strategic STIT formula $\Diamond[\mathcal{H}\ \mathsf{sstit}]\varphi$. (Where $[\mathcal{H}\ \mathsf{sstit}]$ is a *strategic* STIT operator, rather than the STIT operator used in this dissertation.)

[35] Broersen (2011a) mistakenly thinks that this formula captures the intuition of uniform strategies. This is false because an agent may know that it is possible that she ensures that she picks an ace of spades from a face-down deck of cards ($\mathsf{K}_i\Diamond[i\ \mathsf{stit}]\varphi_{\spadesuit A}$ holds), even though she does not know that it is possible for her to knowingly pick an ace of spades (she lacks a uniform strategy to ensure $\varphi_{\spadesuit A}$). We return to this distinction in the next subsection (see Observation 3.1).

[36] This corresponds to the discussion of Herzig and Troquard (2006), more specifically Hypothesis 3 on page 212 and Property 2 on page 214.

in my theorizing is a very weak one that is based on causality. I submit that an agent is able to φ if and only if it is possible that she brings it about that φ. The logical characterization of 'group $\mathcal{H}$ is able to φ' is

$$\Diamond[\mathcal{H}\ \mathsf{stit}]\varphi.^{37}$$

Note that this is importantly different from it being merely possible that φ, that is, $\Diamond\varphi$. It is possible that I lift my arm at the same time as my neighbour blinks in front of the bathroom mirror, yet I certainly lack an ability to do so, in the sense that I cannot guarantee that this will happen. Ability is thus analysed as a nested modal operator: it is possible that I perform an action that guarantees φ.[38]

There is an important distinction between being able to do φ and being able to knowingly do φ. That is, there is a distinction between $\Diamond[i\ \mathsf{stit}]\varphi$ and $\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi$.[39] The distinction I draw is best explained by considering an example. Reconsider the example of Ann, who considers a face-down deck of cards (see also Figure 3.1). Ann is certainly able to pick a spade. That is, there is an action that she can perform which is, as a matter of fact, picking a spade, that is, letting a stand for Ann, $\Diamond[a\ \mathsf{stit}]\varphi_{\spadesuit}$. In contrast, Ann is not able to *knowingly* pick a spade. Indeed, there is no action available that she knows is an instantiation of her picking a spade, i.e. $\Diamond\mathsf{K}_a[a\ \mathsf{stit}]\varphi_{\spadesuit}$ does not hold. Notice that this involves knowledge of her action, which is best viewed as ex interim knowledge. In contrast, in this example Ann knows that she is able to pick a spade, that is, $\mathsf{K}_a\Diamond[a\ \mathsf{stit}]\varphi_{\spadesuit}$. This knowledge concerns static properties: thirteen of these cards are spades, so she is able to pick one of them. She knows this. We thus obtain the following logical (in)validities:

[37] Horty and Belnap (1995) propose a similar characterization, although they rely on deliberative STIT. They show that their characterization straightforwardly connects to Brown's (1988) proposal.

[38] It is certainly not my objective to give a thorough analysis of ability and all its complexities. Some think that knowing-how is merely a kind of ability, and my investigation of knowing-how proves a similar point.

[39] Herzig and Troquard (2006) characterize individual knowing-how roughly by $\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi$, which is equivalent to $\mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi$ in their framework. The same equivalence holds in my framework when (Unif-H) is adopted (see Proposition 3.1).

Observation 3.1 (Properties of Knowledge and Ability)**.** *Let $i \in Ags$ be an agent. Assume (Unif-H). Then the following (schematic) validities and invalidities hold:*

$$\models \Diamond \mathsf{K}_i[i \text{ stit}]\varphi \to \mathsf{K}_i\Diamond[i \text{ stit}]\varphi \qquad \textit{(using } \textbf{\textit{Unif-H}}\textit{)}$$
$$\not\models \mathsf{K}_i\Diamond[i \text{ stit}]\varphi \to \Diamond\mathsf{K}_i[i \text{ stit}]\varphi \qquad \textit{(discussion above)}$$
$$\models \mathsf{K}_i\Diamond[i \text{ stit}]\varphi \to \Diamond[i \text{ stit}]\varphi \qquad \textit{(using } \textbf{\textit{T for}}\ \mathsf{K}_i\textit{)}$$
$$\not\models \Diamond[i \text{ stit}]\varphi \to \mathsf{K}_i\Diamond[i \text{ stit}]\varphi \qquad \textit{(discussion above)}$$

My theory of knowing-how relies on refinements and on action hierarchies (§ 3.3.1). Consider the case in which Bob faces a situation similar to Ann, the only difference being that all cards are face up. A simplification of the scenario is depicted in Figure 3.2. Bob does have the ability to knowingly pick a spade. Indeed Bob knows that he has this ability, that is, letting b stand for Bob, $\mathsf{K}_b\Diamond\mathsf{K}_b[b \text{ stit}]\varphi_{\spadesuit}$. Why? Because there is a refinement, for instance picking the ace of spades, which he knows is a refinement of picking a spade. We say that this is a refinement because every way for him to pick the ace of spades is a way for him to pick a spade. And, furthermore, he knows that he is able to knowingly pick the ace of spades. These two conditions respectively correspond to the two requirements in the following definition of individual know-how:

Definition 3.5 (Individual Know-how)**.** *Let $\mathcal{M}$ be an epistemic STIT model, let $i \in Ags$ be an individual agent, and let $\langle s, h\rangle$ be a dynamic state. Then we say, relative to $\langle s, h\rangle$, that* an individual agent *i* knows how to *φ if and only if there is*[40] *a (proper) refinement ψ of φ for i such that*

1. *She knows that ψ is a (proper) refinement of φ for her, i.e.*

$$\mathcal{M}, \langle s, h\rangle \models \mathsf{K}_i\Box(\mathsf{K}_i[i \text{ stit}]\psi \to \mathsf{K}_i[i \text{ stit}]\varphi);$$

[40] Unfortunately, this existential quantification cannot be captured in the current logic. This means that the concept of individual know-how is not straightforwardly characterizable in the current logic. Nevertheless, this section concludes with a central theorem that shows that individual know-how can be characterized.

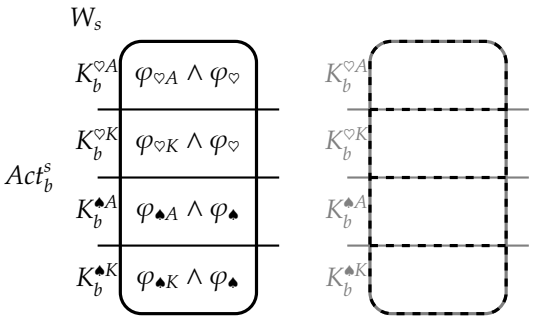


Figure 3.2: *A simplified epistemic STIT model that represents Bob's decision problem: Act_b^s represents the actions available to him; and his information partition equals the set of actions available to him, which means that he can epistemically distinguish the different actions available to him.*

2. *She knows that she is able to knowingly ψ, i.e.*

$$\mathcal{M}, \langle s, h \rangle \vDash \mathsf{K}_i \Diamond \mathsf{K}_i [i \text{ stit}] \psi.$$[41]

It is worth remarking that one could know how to φ without having a proper refinement of φ-ing. An example might clarify this: an agent may know how to ride a bike without being able to explain how she does it. In this case, she knows how to ride a bike despite lacking knowledge of a *proper* refinement of her bike-riding. This is an instance of the above definition where φ expresses 'bike-riding' and $\psi \equiv \varphi$. (Since the formula in condition 1 then turns into a valid

[41] Compare Carr (1979, p. 402 – notation adapted): "On the 'bring about' analysis of action descriptions our ability and knowing how contexts now expand as follows:

> *i* is able to bring it about that φ.
>
> *i* knows how to bring it about that φ.

It transpires on this analysis that the logical form suggested earlier, of sentences about knowing how and ability, was only the apparent form, for such sentences are really instances of one modal construction buried within another. Ability and know how are thus seen to be concepts of third level (like e.g. the 'know' of self-knowledge, as it occurs in sentences like '*i* knows that he knows that φ')." Accordingly, my notion of know-how would be a concept of fifth level.

formula, her knowing-how amounts to condition 2, viz. $\mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi$.) I submit that an agent *implicitly* knows how to φ if and only if she knows how to φ despite *lacking* knowledge of a *proper* refinement.

Alternatively, one could know how to φ by having a proper refinement of φ-ing. An example may help to illustrate this: in the example of the face-up deck of cards, Bob knows that he can pick a spade by picking the ace of spades. In this way, he can explain that he knows how to pick a spade by mentioning that he can do so by picking the ace of spades. Although there may be multiple ways to pick a spade, one witness suffices for him to know how to do so. This is an instance of the above definition where $\varphi \equiv \varphi_{\spadesuit}$ and $\psi \equiv \varphi_{\spadesuit A}$. Moreover, $\varphi_{\spadesuit A}$ is a proper refinement of $\varphi_{\spadesuit}$ for Bob, that is, $\Box([b\ \mathsf{stit}]\varphi_{\spadesuit A} \rightarrow [b\ \mathsf{stit}]\varphi_{\spadesuit})$ holds while its converse does not (see Definition 3.4). I submit that an agent *explicitly* knows how to φ if and only if her knowing-how is witnessed by a *proper refinement* ψ.

The difference between implicit and explicit know-how is best expressed in terms of explainability. When an agent implicitly knows how to φ she is unable to explain how she does it. Conversely, when an agent explicitly knows how to φ she is able to explain how she does it. Although the agent knows how to φ in both cases, the difference is highlighted in my theory by the fact that her knowing-how may either exclude or include a proper refinement.

One may ask whether these notions can be logically characterized. This is the question I now turn to. It might be unsurprising, because of the existential quantification noted in Footnote 40, that neither explicit nor implicit know-how can be characterized:

Theorem 3.2 (Impossibility Results)**.** *Implicit individual know-how and explicit individual know-how cannot be characterized in the current logic.*

That is, given an agent i and a formula φ, there is no *formula ψ such that for every model $\mathcal{M}$ and every dynamic state $\langle s, h\rangle$ it holds that, relative to $\langle s, h\rangle$, agent i implicitly knows how to φ if and only if $\mathcal{M}, \langle s, h\rangle \vDash \psi$. And something similar applies to explicit know-how.*

Given this impossibility result, it seems even more remarkable that individual know-how can be characterized as follows:

Theorem 3.3 (Characterization of Individual Know-how)**.** *Individual know-how is characterized by*

$$\mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi.$$

That is, for every model $\mathcal{M}$, every dynamic state $\langle s, h\rangle$, every individual agent $i \in Ags$, and every formula φ it holds that, relative to $\langle s, h\rangle$, agent i knows how to φ if and only if $\mathcal{M}, \langle s, h\rangle \vDash \mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi$.

This subsection thus culminates in a neat characterization of individual know-how that addresses the raised concerns. This means that the current logical framework is powerful enough to express the concept of individual practical knowledge, despite being unable to express implicit and explicit knowledge. Among other things, this highlights that the complete system provided by Theorem 3.1 can be used to reason about individual practical knowledge.[42]

3.3.3.1 Intellectualism

Ever since Ryle (1949), philosophers have been divided into intellectualists, who think that knowing-how reduces to knowing-that, and anti-intellectualists, who resist such a reduction.[43] Recently, Stanley and Williamson (2001) have argued

[42] Nonetheless, it is unclear whether it is tractable to do so since I have not studied the complexity of either model-checking or satisfiability. For now, I wish to point to the relevant literature on this topic (van der Hoek and Wooldridge, 2003; Herzig and Schwarzentruber, 2008; Payette, 2014).

[43] The debates usually depict two strands of anti-intellectualism: weak, which denies intellectualism, and strong, which embraces the prioricity of knowing-how over knowing-that.

for the intellectualist position, a position that seems to be underpopulated. Theorem 3.3 shows that individual know-how is characterized by $\mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi$, that is, agent i knows that it is possible that she knowingly does φ. I will explain what this characterization establishes in light of this philosophical debate by highlighting three points.

First, it is useful to revisit the premises of the debate and clarify what is typically meant by knowing-how and knowing-that. Jeremy Fantl opens his paper on knowing-how by emphasizing different kinds of knowledge. He writes:

> Contemporary epistemology distinguishes among three kinds of knowledge: *propositional* knowledge, knowledge by *acquaintance*, and *practical* or *procedural* knowledge. Propositional knowledge is what is expressed by sentences relevantly similar to "Alex knows that George W. Bush is the U.S. president". . . . Practical knowledge is what is expressed by sentences relevantly similar to "Callie knows how to ride a bicycle". (Fantl, 2008, p. 451)[44]

I propose to differentiate knowing-how and knowing-that on the basis of the kind of content they take: knowing-that takes a proposition as content, and knowing-how takes an action as content. Given the STIT analysis of action (§ 1.2) and my views on action hierarchies (§ 3.3.1), act descriptions are expressed by locutions such as 'agent i sees to it that φ'.[45] Although I am unsure whether this commonly qualifies as a proposition, such act descriptions are included in knowing-how, but not in knowing-that.[46] So, if one were to accept the statement 'it

[44] One's take on propositional knowledge may affect one's position in the intellectualism debate. For instance, Fantl (2008, pp. 452–453) argues that "[s]trong anti-intellectualism is most plausible if knowing that something is the case is essentially dispositional . . . But if knowing that something is the case is not essentially dispositional, then it seems rather implausible for knowing-that to be reduced to or a species of knowing-how".

[45] Although David Carr so far agrees with my analysis, he eventually writes: "It appears to be the case that knowing how statements require as their objects, descriptions of actions construed in a much more sophisticated way [than descriptions of actions understood merely as instances of bringing about or agent-causation], as, in fact, *intentional actions*" (Carr, 1979, p. 409 – emphasis added).

[46] Compare Glick (2011, p. 413 – notation adapted): "Take, for instance, abilities. For any action of φ-ing, we could map i's ability to φ onto the proposition that i φs, and instead of saying that i is able

is possible that agent i knowingly does φ', expressed by $\Diamond \mathsf{K}_i[i\ \mathsf{stit}]\varphi$, into the realm of propositions then my characterization yields a reduction of knowing-how to knowing-that.

To understand the nature of propositions as used in my formalism it is vital to recall that they are evaluated at *dynamic states*, which include the temporal evolution of the world. Moreover, the valuation assigns to each proposition the set of dynamic states at which it holds. Since the temporal evolution of the world intuitively includes events and processes, propositions can depict various things: static states of affairs, temporal conditions, processes, activities, etc. Accordingly, practical knowledge can be viewed as a species of propositional knowledge.

Second, anti-intellectualists typically argue that knowing-that misses the crucial connection to skills and abilities that is akin to knowing-how.[47] Some have refuted intellectualism because it requires knowing-how to be demonstrable, that is, that one can express or explain its practical knowledge.[48] It is therefore important to note that I distinguished between implicit and explicit practical knowledge to do away with such criticism. Moreover, it should be noted that the characterization of individual know-how retains the connection to skills and abilities despite depicting knowing-how as a species of knowing-that. After all, whenever agent i knows how to φ, on my view, then she has the knowledge and skills to perform an action that guarantees φ.

Third, it may be helpful to address a possible criticism of my characterization of individual know-how, namely that it requires an agent's know-how to be *effective*. It seems that my characterization does not allow for the possibility that

to φ, we could say that i 'ables that he φs'. If we had this linguistic convention, we might note that 'abling' is a relation to a proposition, but of course, by hypothesis, we would be talking about the same thing we actually talk about with ability attributions."

[47] Compare Stanley (2011, p. vii): "If it is surprising that knowledge of a fact can so immediately yield knowledge of how to swim, ride a bicycle, or play a piano, it is only so because of false assumptions about what it is to know a fact. ... There are false assumptions about what it is to *act* on knowledge of facts, there are false assumptions about what it is to have *knowledge* of facts, and there are false assumptions about the *nature* of facts."

[48] Fodor (1968, p. 634) famously writes: "Certain of the anti-intellectualist arguments fail to go through because they confuse knowing that with being able to explain how."

an agent knows how to φ, yet fails to be able to φ here and now. The typical counterexample concerns someone who still knows how to walk even after losing his legs in a tragic accident. I will discuss two ways in which my formalism may be adapted to address this concern. One way of implementing this is by *explicitly* adding some kind of background conditions for the knowing-how. Individual know-how would then consist of four elements: the agent i, the object φ, the refinement ψ, and the background conditions γ.[49] Roughly stated, one would then say that agent i knows how to φ if and only if there is a refinement ψ such that, whenever the background conditions γ hold, agent i knows that she can φ by ψ-ing and she knows that she can knowingly ψ.

Another way is to implement this idea *implicitly*. A particular instance of knowing-how is then investigated by a class of pointed epistemic STIT models, say $\mathcal{C}$, rather than by a single epistemic STIT model $\mathcal{M}$ and a single dynamic state $\langle s, h\rangle$. The modeller should construct this class wisely and, for instance, take the relevant background conditions into account. If, for example, the legless person's knowing how to walk is to be assessed under the background condition that he has legs, then the relevant class of models should depict him as having legs, despite this being contrary to fact. Simply stated, one would then say that agent i knows how to φ if and only if there is a refinement ψ such that, at every pointed epistemic STIT model $\mathcal{M}, \langle s, h\rangle$ in the class $\mathcal{C}$, agent i knows that she can φ by ψ-ing and she knows that she can knowingly ψ.

These two ways of addressing the concern about effectivity are quite common in formal philosophy: whenever one models (aspects of) a particular concept, one is prone to such modelling considerations, especially when the meaning of

[49] In earlier drafts of papers on know-how, my co-author, Jan Broersen, initially insisted that knowing how to φ has these four elements. Adding these background conditions would open up two lines of enquiry. First, what are the correct background conditions for assessing whether agent i knows how to φ? For example, to assert that an agent knows how to ride a bike, should we consider the normal circumstances or perhaps any metaphysically possible circumstances? Second, what is an agent required to know about these background conditions in order for her to know how to φ? For example, to assert that an agent knows how to ride a bike, should the agent know that the normal circumstances obtain?

the concept relies on contextual factors, background conditions, or ceteris paribus clauses. This does not, however, render my modelling exercise to be in vain: my conceptual and formal analysis is aimed at explicating the concept of individual know-how using action hierarchies.

3.4 Collective Know-how

Until now I have been mainly concerned with individual practical knowledge, but I aim to show that these ideas can be fruitfully extended to the collective level. It is important to recall that my theory of collective know-how is designed for situations in which agreement is problematic and communication is hindered. One could therefore view my theory of a group's collective know-how as being somewhat *minimal*, that is, it seems plausible to say that the group's collective know-how expands if these restrictions are lifted. My starting point is that a group collectively knows how to decorate a house together only if there is a (proper) refinement ψ which they collectively know is a way to decorate a house together, and they collectively know that they are jointly able to knowingly ψ.

When does a group know that it is able to knowingly ψ? More specifically, which refinements justify this practical knowledge?[50] Let us consider an example to reveal some intuitions. Suppose that there are two decks of cards face down on the table, one for Chris and one for Dee, and they simultaneously pick a card from their deck. A simplification of this scenario is depicted in Figure 3.3, where c and d stand for Chris and Dee respectively, and $\mathcal{F}$ depicts the group consisting of Chris and Dee.[51] They are certainly able to pick two spades. After all, there is a group action that they can perform which is, as a matter of fact, picking two

[50] Searle (1990, p. 410) writes: "I believe one of the keys to understanding collective intentionality is to see that in general the by and by-means-of relations for achieving the collective goal have to end in individual actions." My discussion of collective know-how, in terms of action hierarchies, relies on a similar intuition.

[51] Note that Chris's and Dee's individual situation is similar to that of Ann, which is represented in Figure 3.1.

spades, that is, $\Diamond[\mathcal{F}\ \mathsf{stit}]\varphi_{\spadesuit\spadesuit}$ holds. Moreover, they commonly know that the only way for them to jointly pick two spades is by each picking a spade, that is, $\mathsf{C}_{\mathcal{F}}\Box([c\ \mathsf{stit}]\varphi_{\spadesuit} \wedge [d\ \mathsf{stit}]\varphi_{\spadesuit} \leftrightarrow [\mathcal{F}\ \mathsf{stit}]\varphi_{\spadesuit\spadesuit})$ holds. However, since neither of them knows how to pick a spade, *neither of them knows how to play her part*. To my understanding, this entails that they do not collectively know how to pick two spades.

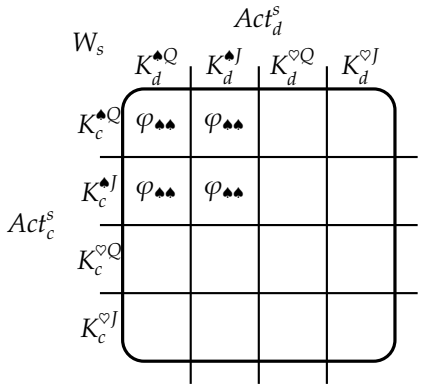


Figure 3.3: *A simplified epistemic STIT model that represents Chris and Dee's joint decision problem: (a) for the example of the face-down decks of cards, each agent's epistemic indistinguishability relation is given by the universal relation on W_s; (b) for the example of the face-up decks of cards, each agent's information partition equals the set of actions available to her.*

At the opposite extreme is an example in which all cards are face up, which is illustrated in Figure 3.3.[52] In this case there is a proper refinement of picking two spades, for instance Chris picking the ace of spades $\psi_{\spadesuit A}$ and Dee picking the jack of spades $\psi_{\spadesuit J}$, and they commonly know that is a way for them to pick two spades. More precisely, we have the following:

[52]Note that Chris's and Dee's individual situation is similar to that of Bob, which is depicted in Figure 3.2.

1. $\mathsf{C}_{\mathcal{F}}\Box([c\ \mathsf{stit}]\psi_{\spadesuit A} \wedge [d\ \mathsf{stit}]\psi_{\spadesuit J} \to [\mathcal{F}\ \mathsf{stit}]\varphi_{\spadesuit\spadesuit})$, that is, Chris and Dee commonly know that Chris picking the ace of spades and Dee picking the jack of spades is a way for them to pick two spades;

2. $\mathsf{C}_{\mathcal{F}}(\mathsf{K}_c\Diamond\mathsf{K}_c[c\ \mathsf{stit}]\psi_{\spadesuit A} \wedge \mathsf{K}_d\Diamond\mathsf{K}_d[d\ \mathsf{stit}]\psi_{\spadesuit J})$, that is, Chris and Dee commonly know that each of them knows how to carry out her part $\psi_{\_}$.[53]

One may object that the 'common knowledge' requirement is too strong for an account of collective knowing-how. Although I think it is worthwhile to explore whether and how this requirement can be weakened, my present worries are complementary to such considerations. After all, an alternative account of collective knowledge could be implemented into my account of collective practical knowledge by augmenting the 'common knowledge' requirement accordingly.

To return to the last example, Chris and Dee both know that Chris's $\psi_{\spadesuit A}$-ing and Dee's $\psi_{\spadesuit J}$-ing jointly constitute a refinement of their picking two spades. In this example, they have exactly the same information. Chris, in particular, knows all the ways in which Door can knowingly pick a spade. Since many everyday collaborations involve information asymmetry, a focus on symmetrical information states is too restrictive for a theory of collective know-how to encompass everyday collaborations.

To investigate collective practical knowledge further, consider a different example, which again concerns two face-down decks of cards. This time, however, there is a box of special glasses. These glasses allow any agent to see through one particular suit. So, if an agent is wearing the hearts glasses, then she is able to see through hearts, that is, this allows her to identify the jack of hearts, the 2 of hearts, etc. Suppose Edo and Fae each take a pair of these glasses and put them on.

[53] Note that in the previous example, where the decks are face down, the first property holds. The problem with that example is that it violates the second property.

As a matter of fact, Edo took the spades glasses and Fae took the hearts glasses. Suppose neither of them knows the other's particular type of special glasses.[54] Do Edo and Fae collectively know how to pick two jacks?

Figure 3.4 represents a simplified version of this complex scenario; let me briefly explain the figure. We let e and f stand for Edo and Fae respectively, and let $\mathcal{G}$ depict the group consisting of Edo and Fae. In this simplified version, the face-down cards only include the jack and queen of hearts and spades, and the box of glasses only contains the pairs of glasses associated with spades or hearts. The STIT model on the left-hand side represents the actions available to each agent, which for instance shows that Edo is able to pick the queen of spades (since $K_e^{\spadesuit Q} \in Act_e^-$). The knowledge of Edo and the knowledge of Fae is represented by the illustration on the right-hand side, which involves four different static states s_1–s_4. Lastly, the figure explicitly states three dynamic worlds u, v, and w. Let us investigate the individual knowing-how in this example. Note that, for instance, at dynamic world w, Edo is wearing the spades glasses and Fae is wearing the hearts glasses. If we concentrate on Edo's epistemic uncertainty, we can see that Edo knows that the actual static state is either s_3 or s_4, not s_1 or s_2. Because Edo cannot distinguish between dynamic worlds w and v, we can observe that, at w, Edo does not knowingly pick the queen of hearts even though he does so unknowingly. Moreover, Edo can distinguish between w and u. In particular, at w, Edo knows how to pick the jack of spades, but he does not know how to pick the jack of hearts.[55] Similarly, we can see that, at w, Fae knows how to pick the jack of hearts, but she does not know how to pick the jack of spades; the opposite holds at v. Finally, because Edo cannot distinguish between w and v, it holds that, at w, he does not know that Fae knows how to pick the jack of hearts.

[54] Although it is a rather artificial example, it will point to a natural concern: even though my companion's implicit practical knowledge is opaque to me, this need not undermine our collective practical knowledge. The artificial example is meant to isolate this particular concern.

[55] The former holds, because one of the cells in his information partition is associated with $K_e^{\spadesuit J}$. The latter holds, because, from Edo's perspective, any dynamic world in $K_e^{\heartsuit J}$ is epistemically indistinguishable from a dynamic world in $K_e^{\heartsuit Q}$.

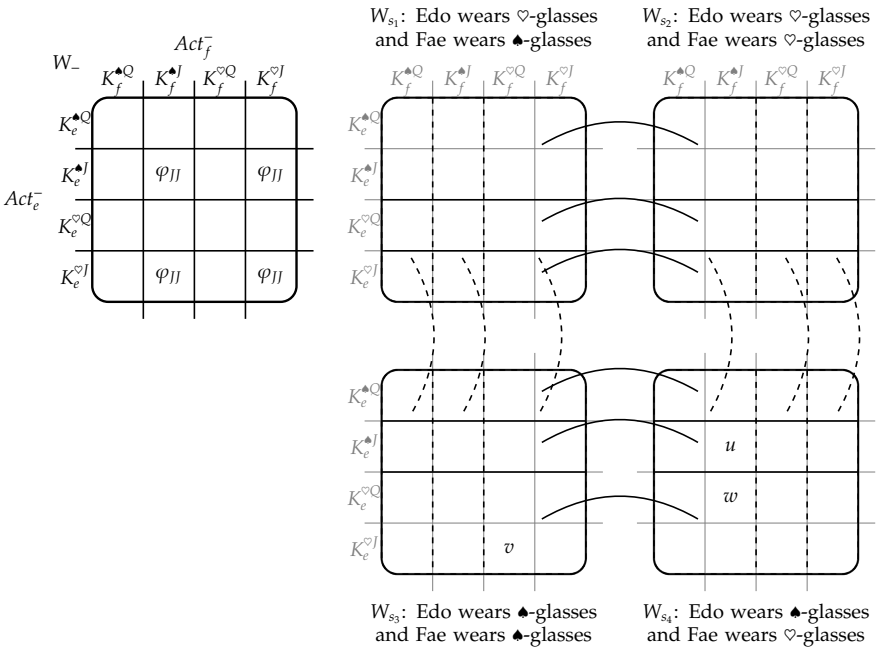


Figure 3.4: *A simplified epistemic STIT model that represents Edo and Fae's joint decision problem: Act_e^s and Act_f^s represent the actions available to Edo and Fae respectively; the solid lines and partitions represent Edo's knowledge, and the dashed lines and partitions represent Fae's knowledge.*

I would like to say that Edo and Fae collectively know how to pick two jacks. It is important to note that we do not allow for communication in this example. Let us start with investigating their individual knowledge and, in particular, with highlighting the asymmetry in their information states. Edo knows that he can pick a jack by knowingly picking the jack of spades, that is, $\mathsf{K}_e\Box(\mathsf{K}_e[e\ \mathsf{stit}]\psi_{\spadesuit J} \to \mathsf{K}_e[e\ \mathsf{stit}]\varphi_J)$ holds. Something symmetric holds for Fae. So, Edo's practical knowledge of φ_J-ing is explicit with respect to his own knowledge. Or, equivalently, *Edo explicitly knows how* to φ_J. Moreover, Edo knows how to pick the jack of spades, that is, $\mathsf{K}_e\Diamond\mathsf{K}_e[e\ \mathsf{stit}]\psi_{\spadesuit J}$ holds. In contrast, Fae does not know that Edo knows this. After all, Fae does not know which type of glasses Edo is

wearing, so Fae does not know whether Edo is able to knowingly pick a jack of spades, diamonds, hearts, or clubs. Hence, there is no specific jack which Fae knows that Edo knows that he can knowingly pick. That is, $\mathsf{K}_f\mathsf{K}_e\Diamond\mathsf{K}_e[e\ \mathsf{stit}]\varphi_J$ holds, but for every $\circ \in \{\heartsuit, \spadesuit, \diamond, \clubsuit\}$ the formula $\mathsf{K}_f\mathsf{K}_e\Diamond\mathsf{K}_e[e\ \mathsf{stit}]\psi_{\circ J}$ does not hold. So Edo's practical knowledge of φ_J-ing is *implicit for Fae*. In sum, the information asymmetry amounts to this: Edo explicitly knows how to φ_J and Edo implicitly knows how to $\psi_{\spadesuit J}$, but Fae does not know that Edo knows how to $\psi_{\spadesuit J}$.

Now, let us investigate their common knowledge in this example. First, note that they commonly know that the only way for them to jointly pick two jacks is by each picking a jack. That is, $\mathsf{C}_{\mathcal{G}}\Box([e\ \mathsf{stit}]\varphi_J \wedge [f\ \mathsf{stit}]\varphi_J \leftrightarrow [\mathcal{G}\ \mathsf{stit}]\varphi_{JJ})$ holds. This means that group $\mathcal{G}$ commonly knows that $[\mathcal{G}\ \mathsf{stit}]\varphi_{JJ}$ divides into parts $[e\ \mathsf{stit}]\varphi_J$ and $[f\ \mathsf{stit}]\varphi_J$. Second, observe that it is commonly known that each of them knows how to play her part. That is, for example, $\mathsf{C}_{\mathcal{G}}\mathsf{K}_e\Diamond\mathsf{K}_e[e\ \mathsf{stit}]\varphi_J$. After all, Fae knows that Edo is able to identify a jack, since she knows that he is wearing a pair of special glasses.

We arrive at a complex picture of the refinements that justify collective know-how in this example:

Observation 3.2. *In the discussed example, we say that group $\mathcal{G}$, consisting of e and f, knows how to φ because the following conditions hold:*

1. $\mathsf{C}_{\mathcal{G}}\Box([e\ \mathsf{stit}]\varphi_J \wedge [f\ \mathsf{stit}]\varphi_J \leftrightarrow [\mathcal{G}\ \mathsf{stit}]\varphi_{JJ})$*: group $\mathcal{G}$ commonly knows that* $[\mathcal{G}\ \mathsf{stit}]\varphi_{JJ}$ *divides into parts* $[e\ \mathsf{stit}]\varphi_J$ *and* $[f\ \mathsf{stit}]\varphi_J$*;*

2. $\mathsf{C}_{\mathcal{G}}\mathsf{K}_e\Diamond\mathsf{K}_e[e\ \mathsf{stit}]\varphi_J$*, and similarly for f: group $\mathcal{G}$ commonly knows that each of its members knows how to carry out her part;*

3. $\mathsf{K}_e\Box(\mathsf{K}_e[e\ \mathsf{stit}]\psi_{\spadesuit J} \rightarrow \mathsf{K}_e[e\ \mathsf{stit}]\varphi_J)$*: Edo knows that there is a proper refinement ψ_- of his part φ_- (and something similar holds for f);*

4. $\mathsf{K}_e\Diamond\mathsf{K}_e[e\ \mathsf{stit}]\psi_{\spadesuit J}$*: Edo knows how to perform his ψ_- (and something similar holds for f).*

The action hierarchy in this definition has *three* levels, the collective action (given by $[\mathcal{G}\ \mathsf{stit}]\varphi_{JJ}$), the division (jointly expressed by $[e\ \mathsf{stit}]\varphi_J$ and $[f\ \mathsf{stit}]\varphi_J$), and the individual refinements (given by $[e\ \mathsf{stit}]\psi_{\spadesuit J}$ and $[f\ \mathsf{stit}]\psi_{\heartsuit J}$).[56] Group $\mathcal{G}$ commonly knows that $[e\ \mathsf{stit}]\varphi_J \wedge [f\ \mathsf{stit}]\varphi_J$ is a refinement of $[\mathcal{G}\ \mathsf{stit}]\varphi_{JJ}$, though not a proper refinement. This means that $\mathcal{G}$ collectively *implicitly* knows how to φ. Each member knows that ψ_- is a proper refinement of her φ_- and knows how to ψ_-. This means that each member *explicitly* knows how to play her own part φ_-. In general this individual practical knowledge need not be explicit. Given the characterization of individual know-how (Theorem 3.3), we can simplify the concept of collective know-how to the following:

Definition 3.6 (Collective Know-how)**.** *To aid readability, I present my definition of collective know-how for a two-agent group $\mathcal{H} = \{i, j\}$. We say that $\mathcal{H}$* knows how to *φ if and only if there are*[57] *φ_i and φ_j such that the following two conditions hold:*

1. $\mathsf{C}_{\mathcal{H}}\Box([i\ \mathsf{stit}]\varphi_i \wedge [j\ \mathsf{stit}]\varphi_j \leftrightarrow [\mathcal{H}\ \mathsf{stit}]\varphi)$*: group $\mathcal{H}$ commonly knows that $[\mathcal{H}\ \mathsf{stit}]\varphi$ divides into parts $[i\ \mathsf{stit}]\varphi_i$ and $[j\ \mathsf{stit}]\varphi_j$;*

   (Common Division)

2. $\mathsf{C}_{\mathcal{H}}\mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi_i$*, and similarly for j: group $\mathcal{H}$ commonly knows that each member knows how to play her part.*

   (Part Know How)

One may ask whether the 'part know how' condition is necessary for collective knowing-how. To see that it is, consider the example where there are two decks of cards face down on the table, one for Chris and one for Dee (see Figure 3.3). Recall that the 'common division' condition is met in this example, but Chris and Dee do

[56] It may be helpful to point out that I will present a hierarchical task network that represents these levels in § 3.5.1.

[57] It is important to note that this existential quantification cannot be captured in the current logic. This means that collective know-how is not straightforwardly characterizable in the current logic. In the remainder of this section it will be shown that, under the assumption that (Unif-H) holds, collective know-how is characterizable.

not collectively know how to pick two spades. We argued that this is due to the fact that Chris and Dee do not commonly know that each knows how to play her part. Or, equivalently, this means that if the 'part know how' condition is violated then the group does not collectively know how to pick two spades. Therefore, the 'part know how' condition is necessary for collective knowing-how.

One may think that the biconditional in the 'common division' condition is too strong. That is, one may think that the left-to-right implication suffices. Can this condition be relaxed accordingly? Suppose we would only require the left-to-right implication: $\mathsf{C}_{\mathcal{H}}\Box([i\ \mathsf{stit}]\varphi_i \wedge [j\ \mathsf{stit}]\varphi_j \rightarrow [\mathcal{H}\ \mathsf{stit}]\varphi)$. It is useful to briefly consider a standard coordination game: to stick with the chapter's type of examples, consider the example where there are two decks of cards face up on the table, one for Chris and one for Dee. Do Chris and Dee know how to pick an identical face card? (Recall that we do not allow for any communication.)[58] Figure 3.5 presents a simplified STIT model of this example, where φ_{id} expresses that they pick an identical face card.[59] Note that the left-to-right implication holds since Chris and Dee commonly know that if each would pick a jack, then they would be jointly picking an identical face card – formally, $\mathsf{C}_{\mathcal{H}}\Box([c\ \mathsf{stit}]\varphi_J \wedge [d\ \mathsf{stit}]\varphi_J \rightarrow [\mathcal{H}\ \mathsf{stit}]\varphi_{id})$ holds. Moreover, note that they commonly know that each knows how to pick a jack, which means that something comparable to the 'part know how' condition holds. In this example, the group faces a coordination problem that typically cannot be solved without prior communication. The members are unable to effectively coordinate their individual actions.[60] Therefore, the group does not know how to pick an identical face card on my view. This coordination

[58] The game resembles the driving game discussed by Lewis (1969, pp. 6, 44–45 – see also § 1.1).

[59] Note that Chris's and Dee's individual situation is similar to that of Bob, which is represented in Figure 3.2.

[60] Although Chris and Dee are not able to solve their coordination game without prior communication, people are often quite effective at solving such coordination problems. Game theorists typically argue that concepts such as *salience*, *focal points*, and *framing* are key to understanding the coordination abilities of non-communicating agents (see Schelling, 1960; Lewis, 1969; Gauthier, 1975; Sugden, 1993, 1995, 2003; Bacharach, 2006). Our notion of collective know-how is best viewed as the ability of a group of non-communicating agents that does not rely on such additional concepts.

problem hence indicates that the one-way implication is insufficient for a theory of collective know-how that applies to cases where agreement is problematic and communication is hindered.[61]

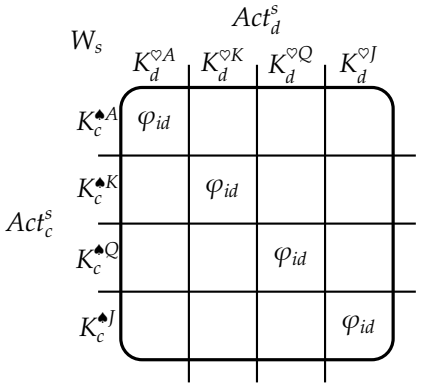


Figure 3.5: *A simplified epistemic STIT model that represents Chris and Dee's joint coordination problem: Act_c^s and Act_d^s represent the actions available to Chris and Dee respectively; and each agent's information partitions equals the set of actions available to her.*

It is important to note that this indicates a divergence from individual practical knowledge. In the example of the face-up deck of cards (Figure 3.2), Bob knows how to pick a face card, even though there are multiple ways of realizing this. After all, for the individual case, we were satisfied with the formula $\mathsf{K}_i\Box(\mathsf{K}_i[i\ \mathsf{stit}]\psi \rightarrow \mathsf{K}_i[i\ \mathsf{stit}]\varphi)$ (see Definition 3.5). The divergence is due to the fact that an individual agent is able to decide to pick the jack of spades and then execute this plan. In the absence of communication, a group may not be able to do this. (At the end of this section I will discuss how communication and agreements may help improve a group's abilities.)

[61] It is unclear how ATL-based paradigms deal with this difficulty (Ågotnes et al., 2015; Jamroga and Ågotnes, 2007; with the possible exception of (Hawke, 2017)). I believe that this lacuna is an artefact of the fact that these formalisms only mention two levels: the goal φ and a (maximally specific) strategy s.

Nonetheless, the problem with a one-way implication in the collective case is not that the collective act can be realized in *multiple* ways, although the problem does not arise if there is a unique way to realize the collective act. After all, in the example of two face-up decks of cards, Chris and Dee collectively know how to pick two, possibly non-identical, face cards, which can be realized in multiple ways. The crucial condition is that when each member of the group plays her alleged part then they jointly succeed in realizing the collective act. The difficulty emphasized in the coordination problem is that playing their alleged parts may not be enough to ensure that they pick an identical face card.[62]

This discussion brings us to the following question: what is a member's *part* in a collective action? There seem to be at least two ways to describe this.[63] Let $[\mathcal{H}\ \mathsf{stit}]\varphi$ be a collective action. First, one may require that the collective action reduces to individual parts φ_i, one for each $i \in \mathcal{H}$. That is, $\mathcal{H}$ jointly brings about φ if and only if each member i plays her part φ_i – this is formally expressed by $\Box([\mathcal{H}\ \mathsf{stit}]\varphi \leftrightarrow \bigwedge_{i\in\mathcal{H}}[i\ \mathsf{stit}]\varphi_i)$. Second, one may simply say that a member i plays her part in the collective action if and only if she performs her component individual action of an instance of this collective action, that is, if and only if $\langle i\ \mathsf{stit}\rangle[\mathcal{H}\ \mathsf{stit}]\varphi$. Or, equivalently, playing one's part means not obstructing the collective act. I prove that, under the assumption that (Unif-H) holds, when $\mathcal{H}$ collectively knows how to φ, these readings are equivalent:[64]

Theorem 3.4. *Let $\mathcal{M}$ be an epistemic STIT model, and let $\langle s, h\rangle$ be a dynamic state. Assume that $\mathcal{M}$ satisfies (Unif-H). Suppose $\mathcal{H}$ collectively knows how to φ, as witnessed by $(\varphi_i)_{i\in\mathcal{H}}$. Then*

$$\mathcal{M}, \langle s, h\rangle \vDash \mathsf{C}_{\mathcal{H}}\Box \bigwedge_{i\in\mathcal{H}}([i\ \mathsf{stit}]\varphi_i \leftrightarrow \langle i\ \mathsf{stit}\rangle[\mathcal{H}\ \mathsf{stit}]\varphi), \qquad (*)$$

$$\mathcal{M}, \langle s, h\rangle \vDash \mathsf{C}_{\mathcal{H}}\Box \bigwedge_{i\in\mathcal{H}}(\mathsf{K}_i[i\ \mathsf{stit}]\varphi_i \leftrightarrow \mathsf{K}_i\langle i\ \mathsf{stit}\rangle[\mathcal{H}\ \mathsf{stit}]\varphi). \qquad (**)$$

[62] This observation aligns with my work in Chapter 2. The connections will be emphasized by Observation 3.3 and the associated discussion.

[63] These ideas originate from my previous work (Tamminga and Duijf, 2017); see also Chapter 2 and, more specifically, §§ 2.3–2.4. In Chapter 4 I develop a more demanding account of participation.

[64] To be precise, (Unif-H) is used to prove (**) below.

Given these equivalences, it can be proven that collective know-how can be characterized. The final picture of the refinements that justify collective know-how can be simplified accordingly:

Corollary 2 (Characterization of Collective Know-how)**.** *To aid readability, I present this corollary on the characterization of collective know-how for a two-agent group* $\mathcal{H} = \{i, j\}$*. Assume (Unif-H). Then group* $\mathcal{H}$ *knows how to* φ *if and only if the following two conditions hold:*

1. $\mathsf{C}_{\mathcal{H}}\Box(\langle i \text{ stit}\rangle[\mathcal{H} \text{ stit}]\varphi \wedge \langle j \text{ stit}\rangle[\mathcal{H} \text{ stit}]\varphi \leftrightarrow [\mathcal{H} \text{ stit}]\varphi)$*: group* $\mathcal{H}$ *commonly knows that* $[\mathcal{H} \text{ stit}]\varphi$ *divides into parts;*

   (Effectivity)

2. $\mathsf{C}_{\mathcal{H}}\mathsf{K}_i\Diamond\mathsf{K}_i\langle i \text{ stit}\rangle[\mathcal{H} \text{ stit}]\varphi$*, and similarly for* j*:* $\mathcal{H}$ *commonly knows that each of them knows how to carry out her part.*

   (Part Know-how)

It is important to highlight that the problematic existential quantification of my initial definition of collective know-how has disappeared (see Footnote 57). The previous theorem establishes that when a group commonly knows that the group action $[\mathcal{H} \text{ stit}]\varphi$ divides into parts, then these parts can be simply expressed by $\langle i \text{ stit}\rangle[\mathcal{H} \text{ stit}]\varphi$, and no reference to the initial division, that is, $(\varphi_i)_{i\in\mathcal{H}}$, is needed.

The corollary highlights that my epistemic STIT logic is able to characterize collective know-how. By providing the formal framework, I contribute to theories of collective know-how by clarifying and specifying particular elements needed to overcome the discussed conceptual difficulties. Moreover, my current logic is able to express the complex notion of collective know-how that I defended, despite it being unable to express the distinction between explicit and implicit practical knowledge.

It may be useful to explain why the first condition has been renamed 'effectivity' when it was previously called 'common division'. In light of Theorem 3.4,

given collective know-how, each member's part is clear: it is characterized by $\langle i \text{ stit}\rangle[\mathcal{H} \text{ stit}]\varphi$. This means that the common division is also evident. This observation suggests that the first condition means that the group commonly knows that if each member carries out her alleged part, then they will jointly succeed in guaranteeing φ. Because it is important that this common distribution is effective, it seems apt to rename the condition 'effectivity'. Under this interpretation, the first condition does not require a common division but it does require common knowledge of *effectivity*: the group commonly knows that its members can effectively ensure that together they guarantee φ by each member playing her part.

The characterization of collective know-how can be used to show that there are several ways in which a group may lack practical knowledge. It naturally reveals ways in which communication and agreements may help to improve a group's abilities; I will briefly discuss four of these opportunities. First, the group may lack an effective common division. This means that there are multiple ways in which the group may achieve its goal φ and that these constitute a coordination problem. In the absence of communication, a group fails to collectively know how to φ. In Chapter 2 I studied cases in which communication is possible and agreement unproblematic. We saw that in such cases there are good plans and bad plans. One of the critical conditions of a good plan is that it is interchangeable (recall §§ 2.3–2.4, more specifically, Definition 2.7). It can be proven that the effectivity condition relates to interchangeability:

Observation 3.3 (Interchangeability & Effectivity)**.** *Let S be a game model, let $\mathcal{G}$ be a group of agents, and let φ be a formula. The following are equivalent:*

1. $S \vDash \Box([\mathcal{G} \text{ stit}]\varphi \leftrightarrow \bigwedge_{i\in\mathcal{G}}\langle i \text{ stit}\rangle[\mathcal{G} \text{ stit}]\varphi)$*;*

2. $P_{\mathcal{G}} := \{a_{\mathcal{G}} \in A_{\mathcal{G}} \mid S, a \vDash [\mathcal{G} \text{ stit}]\varphi\}$ *is interchangeable.*

To solve its coordination problem the group should agree on a good group plan to achieve its goal. As argued in § 2.3, a group plan, once agreed upon, coordinates the individual actions of the members of the group: it allows them to act *simultaneously* and *unconditionally*, in the full belief that every group member acts according to the plan. An interchangeable and optimal group plan *guarantees* that if every group member acts according to the plan, then the group itself fulfils its collective obligation (see the end of § 2.4.2 for an informal proof of this conditional and the role of interchangeability). The above correspondence implies that when group $\mathcal{G}$ collectively knows how to φ then the group plan expressed by $[\mathcal{G}\ \mathsf{stit}]\varphi$ is interchangeable. More specifically, collective know-how requires that it be commonly known that the corresponding group plan is interchangeable. So, when the group lacks an effective common division, it can solve this by agreeing on an effective and interchangeable group plan.

Second, the group may lack *common knowledge* of the fact that the division is effective. Or, equivalently, the collective action $[\mathcal{G}\ \mathsf{stit}]\varphi$ might, as a matter of fact, depict an interchangeable plan, yet this fact may not be commonly known. Besides agreeing on a good group plan, to retain common knowledge of effectivity the group could refine its public knowledge through communication. By sharing their private information the group members may learn that the effectivity condition is satisfied. In this way the group retains common knowledge of effectivity without agreeing on a good group plan.

Third, one or more group members may lack the relevant individual know-how. This means that the 'part know-how' condition is violated. This failure can be overcome by improving a group member's individual know-how. While the previous break-downs required communication or agreement, this failure could be overcome without any communication, for instance by practising in solitude.

Communication and agreement may, nonetheless, help to address this issue. The defective group member may learn from others, through communication, how to carry out her part.

Fourth, all members might have the relevant individual know-how, yet together they may lack common knowledge that this is the case (even if the effectivity condition is satisfied). This defect is typically resolved by publicly announcing one's capacities and individual know-how. In this way the group retains common knowledge of part know-how without improving any member's individual know-how.

A study of when and how agents can effectively jointly communicate to adopt these improvement mechanisms has to be left for another occasion.

3.5 Related Artificial Intelligence Research

I have presented a conceptual analysis of collective know-how using an epistemic extension of STIT theory. The key elements of my conceptual analysis are refinements, action hierarchies, knowledge, knowingly doing, and abilities. In short, my theory of a group's collectively knowing how to φ amounts to this: (1) they commonly know that φ divides into members' parts; and (2) they commonly know that each of them knows how to play her part. In this section I relate my work to research in artificial intelligence.

3.5.1 AI Planning and Hierarchical Planning

One of the subfields of artificial intelligence is concerned with automated planning (see the textbook by Ghallab, Nau, and Traverso, 2004). My ideas on action hierarchies naturally relate to one of the subfields of automated planning, viz. *hierarchical* planning.[65] Wikipedia contributors write:

[65]Other subfields include conformant planning, (partially observable) Markov decision processes, and multi-agent planning.

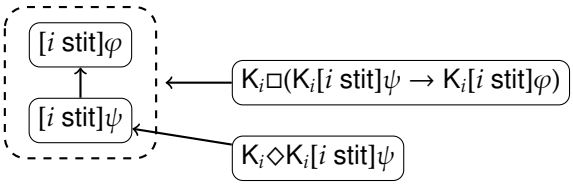
> Planning problems are specified in the *hierarchical task network* approach by providing a set of tasks, which can be:
>
> 1. primitive tasks, which roughly correspond to the actions of STRIPS;
> 2. compound tasks, which can be seen as composed of a set of simpler tasks;
> 3. goal tasks, which roughly correspond to the goals of STRIPS, but are more general.[66]
>
> ...Constraints among tasks are expressed in the form of networks, called task networks. A task network is a set of tasks and constraints among them. Such a network can be used as the precondition for another compound or goal task to be feasible. (Wikipedia contributors, 2017, emphasis added)

Primitive tasks are actions that can be executed, compound tasks are composed of sequences of actions, and goal tasks are tasks that are done to satisfy a condition. A STIT-formula $[i \ \mathsf{stit}]\varphi$ can best be compared to a goal task that needs to be done for agent i to satisfy φ. My STIT framework therefore naturally relates to goal-based hierarchical networks.

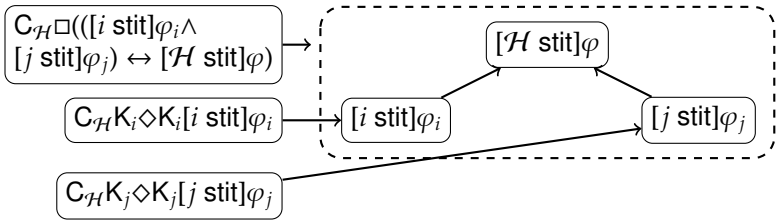
For individual know-how, the constraints of such a goal-based hierarchical network that I am mostly interested in are refinements (§ 3.3.1 and § 3.3.3). That is, I am mostly interested in constraints between goal tasks $[i \ \mathsf{stit}]\varphi$ and $[i \ \mathsf{stit}]\psi$ that are of the following sort: guaranteeing ψ is a way to achieve φ. More specifically, to possess the relevant individual know-how, the agent is required to *know* that guaranteeing ψ is a way to achieve φ. So a goal-based hierarchical

[66] STRIPS, short for STanford Research Institute Problem Solver, is a problem-solving program that has been designed "to find some composition of operators that transforms a given initial world model into one that satisfies some stated goal condition" (Fikes and Nilsson, 1971). One particular formalism that has been widely used to represent hierarchical task networks is TAEMS (Decker, 1995; Horling et al., 1999).

network that represents individual know-how consists of a network of goal tasks such that a link between two goal tasks signifies that the individual agent knows that the former goal task is a refinement of the latter goal task. Figure 3.6a depicts a goal-based hierarchical network for individual know-how.



(a) *A goal-based hierarchical task network depicting individual know-how.*



(b) *A multi-agent goal-based hierarchical task network depicting collective know-how.*

Figure 3.6

My discussions of collective know-how highlighted that a corresponding multi-agent goal-based hierarchical network could consist of several layers: (i) a collective goal $[\mathcal{H}\ \text{stit}]\varphi$, (ii) a division $([i\ \text{stit}]\varphi_i)_{i\in\mathcal{H}}$, and (iii) a collection of refinements $[i\ \text{stit}]\psi_i$ (§ 3.4). The constraints between these layers vary: it should be commonly known that $[\mathcal{H}\ \text{stit}]\varphi$ divides into $([i\ \text{stit}]\varphi_i)_{i\in\mathcal{H}}$, each member i should know that $[i\ \text{stit}]\psi_i$ refines $[i\ \text{stit}]\varphi_i$, and each member i should be able to execute $[i\ \text{stit}]\psi_i$. This means that collective know-how may be represented by such a complex multi-agent goal-based hierarchical network. Figure 3.6b depicts a simplified two-agent goal-based hierarchical task network for collective know-how, where the goal-based hierarchical task networks for the members' individual know-how are suppressed.

3.5.2 Agent-based Artificial Intelligence

In artificial intelligence research there is a strand of literature that recognizes and argues for a new paradigm: an *agent-oriented* approach, rather than an object-oriented approach (Shoham, 1993; Russell and Norvig, 1995; Wooldridge and Jennings, 1995; Yu, 2001). Michael Wooldridge writes:

> But agents are not simply objects by another name. This is because an agent is a rational decision making system: we require an agent to be capable of reactive and pro-active behaviour, and of interleaving these types of behaviour as the situation demands. The object-oriented research community has nothing whatsoever to say about building systems that are capable of this kind of behaviour. In contrast, the design of such systems is a fundamental research topic in the intelligent agents community. (Wooldridge, 1997, pp. 26–27)[67]

One might be sceptical about artificial intelligence ever reaching a point equivalent to human-level agency, but recent developments in artificial intelligence have shown that it may well surpass human intelligence (at least in some domains).[68] It may therefore make sense to treat such artificial intelligent systems as agents, rather than objects. Regardless of these recent breakthroughs, an agent-based approach may help us *design* complex systems. Nicholas Jennings argues for two central claims:

> **The Adequacy Hypothesis.** Agent-oriented approaches can significantly enhance our ability to model, design and build complex, distributed software systems.

[67] Compare Shoham (1993, p. 52): "Most often, when people in AI use the term 'agent', they refer to an entity that functions continuously and autonomously in an environment in which other processes take place and other agents exist. This is perhaps the only property that is assumed uniformly by those in AI who use the term. The sense of 'autonomy' is not precise, but the term is taken to mean that the agents' activities do not require constant human guidance or intervention."

[68] In recent years, artificial intelligence has surpassed humans in *Go* (Metz, 2016), something that was long thought practically impossible, in *Jeopardy!* (Markoff, 2011), and in *Poker* (Metz, 2017), which is relevant because it involves uncertainty.

> **The Establishment Hypothesis.** As well as being suitable for designing and building complex systems, the agent-oriented approach will succeed as a mainstream software engineering paradigm. (Jennings, 2000, p. 278)

Issues concerning cooperation and competition are central to multi-agent systems. It is therefore vital for the agent-based approach to artificial intelligence to study action and knowledge in a precise and formal way. To highlight the impact of my characterization and discussion of collective know-how, I will briefly mention some implications for this agent-based approach.

First, when designing a multi-agent system, one typically uses model-checking techniques to check whether some property Q holds for the system (Alur et al., 2002; van der Hoek and Wooldridge, 2003). Perhaps you want to check whether adversaries collectively know how to break a user's encryption. To check for this, one constructs a model of the multi-agent system and then verifies whether the adversaries possess the relevant collective know-how. An adequate account of collective know-how is thus needed for such model-checking techniques, which is exactly what this chapter is about: analysing the concept of collective know-how and accurately formalizing it.

Second, intelligent agents inside a multi-agent system are typically modelled using mental constructs such as beliefs, desires, and intentions.[69] This has given rise to several BDI-based[70] theories of intelligent agents (Cohen and Levesque, 1990; Rao and Georgeff, 1991, to cite a few). From the perspective of the members of a group of intelligent agents $\mathcal{H}$, it seems that an important condition for adopting a collective goal φ is that they collectively know how to φ. If a group of agents

[69] Shoham (1993, p. 52) writes: "An agent is an entity whose state is viewed as consisting of mental components such as beliefs, capabilities, choices, and commitments. These components are defined in a precise fashion, and stand in rough correspondence to their common sense counterparts. In this view, therefore, agenthood is in the mind of the programmer: What makes any hardware or software component an agent is precisely the fact that one has chosen to analyse and control it in these mental terms."

[70] 'BDI' stands for 'beliefs, desires, and intentions'.

does not collectively know how to φ, then what do they gain from adopting φ as their collective goal? Even if collectively knowing how to φ is not a necessary condition for adopting the collective goal φ, it would be an important property. My characterization of collective know-how allows individual intelligent agents to reason about their joint capabilities (for instance, using the complete logical system of Theorem 3.1).

3.5.3 Logics of Knowledge and Action

I am certainly not the first person to present a formal theory of knowledge and action. The literature on knowledge typically distinguishes between three types of group knowledge: everybody in $\mathcal{H}$ knows φ, typically expressed by $\mathsf{E}_{\mathcal{H}}\varphi$, distributed knowledge that φ, expressed by $\mathsf{D}_{\mathcal{H}}\varphi$, and common knowledge that φ, expressed by $\mathsf{C}_{\mathcal{H}}\varphi$ (Meyer and van der Hoek, 1995; Fagin et al., 2003). To situate my theory of collective know-how within the existing literature on logics of action and knowledge, next I discuss two dominant views in light of this standard distinction.

First, Herzig and Troquard (2006, p. 210) "assume that a coalition can ensure φ if by *sharing their knowledge* and acting together they can ensure φ". This is why their notion of collective know-how relies on *distributed* knowledge. This assumption is problematic for two reasons. First, taking distributed knowledge as a foundation implies that a coalition collectively knows how to φ if there is a way for its members to communicate such that they know how to φ afterwards. But the whole point of collective know-how is that its members are required to know this *before* communicating. After all, distributive knowledge merely marks the possibility, or potentiality, that the know-how will become actual, not that a coalition collectively knows how to φ here and now. Second, I agree with Ågotnes and Wáng (2016, p. 31) that it is a mistake to think that "something is distributed

knowledge in a group if the agents in the group could get to know it after some (perhaps unlimited) communications between them". So, it is even doubtful that distributed knowledge is the correct formalization of "sharing our knowledge".

Second, Ågotnes, Goranko, Jamroga, and Wooldridge (2015, p. 575–576) analyse the notion of "knowing how to play", for the collective case, by appealing to "several different 'modes' in which they can know the right uniform strategy", referring to mutual, distributed, or common knowledge. My conceptual analysis gives two reasons why this categorization is not refined enough. First, when a group collectively knows how to φ then its members commonly know that this divides into members' parts, but the group need not commonly know the exact uniform strategy that each member employs. So, it is not common knowledge of the right strategy that is essential, but common knowledge of *an effective division*. Second, although collectively knowing how to φ implies distributively knowing the right uniform strategy, the converse does not hold. (A case in which I know your part and you know mine proves this point.) So, although collective know-how is a case of distributive knowledge, it is of a particular kind: each member is required to know how to carry out *her own part*. These observations jointly show that neither common nor distributed knowledge captures the intricacies of collective know-how.

In conclusion, justified by the discussions of various examples, I have proposed a characterization of collective know-how that contains elements that have not been recognized in the existing literature. First, my characterization uses action hierarchies. My final definition (Definition 3.6) refers to *three* levels in such action hierarchies: a collective goal, a refinement that specifies each members' part, and a set of individual refinements. In contrast, existing theories typically rely on two levels: the goal, and the strategy or action type. So, although collective practical knowledge has been investigated before in other formal frameworks, its interplay with action hierarchies has been given little weight or attention. I have

shown that this emphasis is useful for the individual case and essential for the collective case. In the individual case, it gives a principled distinction between implicit and explicit practical knowledge. In the collective case, it is needed to distinguish between the members' individual practical knowledge and the collective practical knowledge. Second, existing accounts typically rely on either distributed or common knowledge. However, I argued that neither common nor distributed knowledge captures the intricacies of collective know-how.

Although my study adds to the existing literature, it still leaves interesting questions unanswered. First, what is the logic of collective know-how? My formalism can be used to uncover logical properties of collective know-how. For instance, in the next section I will show that the logic of collective know-how is not monotonic (cf. Wang, forthcoming). Second, although Theorem 3.1 provides a finite axiomatization of epistemic STIT logic, it remains an open question whether this logic and the corresponding model-checking problem are decidable (important pointers include the work by van der Hoek and Wooldridge (2003); Herzig and Schwarzentruber (2008); and Payette (2014)).

3.6 Conclusion

The overall aim of this dissertation is to shed light on collective responsibility problems. As highlighted in § 2.6, a theory of obligations, both individual and collective, is crucial for a theory of responsibility. In this conclusion I aim to briefly investigate individual and collective obligations and responsibility from the perspective of my theory of practical knowledge, as discussed in this chapter.

It is natural to distinguish between objective obligations and subjective obligations. Subjective obligations take into account the epistemic state of the agent.[71] The distinction is based on the intuition that an agent can only be subjectively obliged to do something if she possesses the relevant knowledge. Inspired by the

[71] These should not be conflated with obligations regarding the epistemic state of the agent, which are guided by epistemic norms, such as belief consistency.

dictum that 'ought implies can', my epistemic account of individual and collective ability gives rise to the requirement that if an agent has a subjective obligation to do φ then she has to know how to φ.[72] So practical knowledge is essential for subjective obligations. In this section I will explore the implications of my account of individual and collective know-how without developing a detailed account of subjective obligations.

To showcase how my theory of individual know-how may shed light on individual subjective obligations, let us study a well-known example: the miners' problem. Niko Kolodny and John MacFarlane write:

> Ten miners are trapped either in shaft A or in shaft B, but we do not know which. Flood waters threaten to flood the shafts. We have enough sandbags to block one shaft, but not both. If we block one shaft, all the water will go into the other shaft, killing any miners inside it. If we block neither shaft, both shafts will fill halfway with water, and just one miner, the lowest in the shaft, will be killed. (Kolodny and MacFarlane, 2010)[73]

It is often assumed that the miners' problem warrants the following outcomes of deliberation:

1. We ought to block neither shaft.

[72] The 'ought implies can' principle has often been ascribed to Immanuel Kant, although this attribution is debated. Vranas (2007) gives a defence of the principle, and writes: "I understand the claim that an agent can do something as the claim that the agent has both the ability and the opportunity to do the thing. The agent has the ability to do the thing in the sense of having the requisite skills, physical capacities, and knowledge" (p. 169).

[73] Kolodny and MacFarlane (2010) take the example from Parfit (1988) who in turn credits Regan (1980, p. 265).

Regan (1980, p. 265) can be taken to say that if we assign equal subjective probabilities to the fact that the miners are trapped in shaft A rather than B, then act-utilitarianism requires us to block neither shaft, even though blocking neither shaft cannot possibly be the best act in the circumstances, given the actual location of the miners. Wherever the miners are located, there is an act which is preferable to blocking neither shaft. Regan (1980, p. 265 – terminology adapted) concludes: "Still a reasonable approach to the [miners'] problem requires [us] to abandon all hope of producing the best consequences possible."

2. If the miners are in shaft A, we ought to block shaft A.

3. If the miners are in shaft B, we ought to block shaft B.

4. Either the miners are in shaft A or they are in shaft B.

5. Either we ought to block shaft A or we ought to block shaft B.[74]

I will show how my account of individual practical knowledge, aided by the dictum of ought implies can, may help to dispel the apparent paradox in the miners' problem.[75] Figure 3.7 depicts the miners' problem in an epistemic STIT model. The 'ought implies can' dictum entails that if an agent cannot do φ then she cannot possibly be obliged to do φ. It is not my aim to provide a theory of what the particular subjective obligation *is* in the miners' problem; rather, it is to investigate the practical knowledge of the individual agent in this model and then map a space for the *possible* subjective obligations.

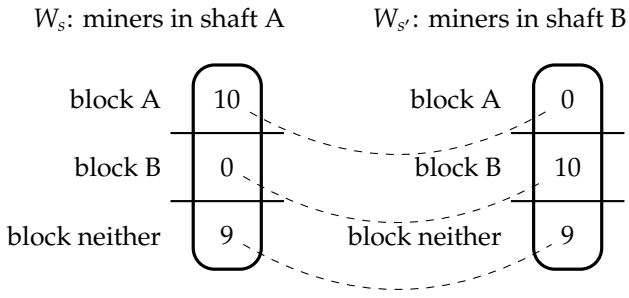


Figure 3.7: *An epistemic STIT model of the miners' problem: the numbers represent the number of miners saved; the epistemic indistinguishability relation is represented by the dashed lines, that is, for instance, the agent cannot distinguish between blocking shaft A in s and blocking shaft A in s'.*

[74] Although the obligations expressed in 1–5 take 'we' as a subject, these obligations depict a general individual agent rather than a collective agent.

[75] Kolodny and MacFarlane (2010), in contrast, argue that "the best way to resolve the paradox is to give a semantics for deontic modals and indicative conditionals that lets us see how the argument can be invalid even with its obvious logical form".

To begin with, note that the agent obviously knows how to block shaft A, how to block shaft B, and how to block neither shaft. In contrast, the problem with assumptions 2 and 3 above is that she does not know how to simultaneously perform the corresponding conditional actions. Let me explain why. Given 4, the only way for her to knowingly ensure that she blocks shaft A if the miners are in shaft A is by knowingly blocking shaft A. Something similar holds for shaft B. Formally, this observation translates to the formula $\mathsf{K}_i\Box(\mathsf{K}_i[i\ \mathsf{stit}](p_A \rightarrow [i\ \mathsf{stit}]A) \rightarrow \mathsf{K}_i[i\ \mathsf{stit}]A)$, where we let p_A stand for 'miners are in shaft A' and let $[i\ \mathsf{stit}]A$ stand for 'i blocks shaft A'.[76] Therefore, the only way for her to knowingly simultaneously (i) block shaft A if the miners are in shaft A and (ii) block shaft B if the miners are in shaft B is by knowingly blocking both shaft A and shaft B, which is impossible. She therefore does not know how to simultaneously perform the conditional actions that correspond to assumptions 2 and 3. Hence, assumptions 2 and 3 jointly lead to a conflict in the required practical knowledge.

In addition, notice that the agent knows that she is able to knowingly save nine miners, namely by knowingly blocking neither shaft. In contrast, she knows that knowingly blocking shaft A is a refinement of both (a) risking that all miners are killed and (b) taking the chance that all miners are saved. Something similar holds for shaft B. In particular, this means that there is no subjective normative difference between knowingly blocking shaft A and knowingly blocking shaft B. Hence, the subjective obligation is either: (i) to knowingly block neither shaft, (ii) to knowingly block one of the shafts, or (iii) to knowingly do anything.[77] Accordingly, assumption 5 is plainly false if read as an exclusive disjunction.

Next, we move on to a study of subjective collective obligations. For the *logical* study of subjective collective obligations, it is crucial to note that collective

[76] A similar observation inspired my study of conditional strategies (Duijf, 2015; Duijf and Broersen, 2016).

[77] Ultimately, it seems plausible that the subjective obligations depend on the morally right risk attitude: (I) if she ought to be risk-averse, then she could be obliged to knowingly block neither shaft, (II) if she ought to be risk-seeking, then she could be obliged to knowingly block one of the shafts, or (III) if she ought to be risk-neutral, then she could be obliged to knowingly do anything.

know-how, unlike individual know-how, is not monotonic. That is, letting $\mathsf{CH}_{\mathcal{H}}\varphi$ stand for '$\mathcal{H}$ collectively knows how to φ', it might be that $\vDash \varphi \to \psi$ while $\nvDash \mathsf{CH}_{\mathcal{H}}\varphi \to \mathsf{CH}_{\mathcal{H}}\psi$. After all, recall the discussions of the coordination game (Figure 3.5): when Chris and Dee independently pick cards from face-up decks of cards, they do not collectively know how to jointly pick an identical face card, despite collectively knowing how to pick two jacks. Since picking two jacks is a way for them to pick an identical face card, this example shows that collective know-how is not monotonic: in the model $\mathcal{M}$ of Figure 3.5, it is the case that $\mathcal{M} \vDash \varphi_{JJ} \to \varphi_{id}$ while $\mathcal{H}$ collectively knows how to φ_{JJ} despite not collectively knowing how to φ_{id} . The intuitive reason is that it may be clear how a specific group action divides into members' parts, while this may be obscured by a more coarse-grained group action. Under the assumption that 'ought implies can', this entails that it is plausible that subjective collective obligations are also not monotonic. That is, letting $\mathsf{SO}_{\mathcal{H}}\varphi$ stand for '$\mathcal{H}$ subjectively collectively ought to φ', it might be that $\vDash \varphi \to \psi$ while $\nvDash \mathsf{SO}_{\mathcal{H}}\varphi \to \mathsf{SO}_{\mathcal{H}}\psi$. The logic of subjective collective obligations therefore has to be non-standard.

What, if anything, does my theory of practical knowledge imply for the relation between collective and individual blameworthiness, that is, backward-looking moral responsibility? I take it that the concept of backward-looking moral responsibility supports the claim that not fulfilling a subjective obligation is a necessary condition for being individually blameworthy: if an individual agent is individually blameworthy, then she has failed to fulfil a subjective obligation. The converse does not hold, because she might have a plausible excuse for not doing what she ought to do. Analogously, I submit that not fulfilling a subjective collective obligation is a necessary, but not a sufficient, condition for being collectively blameworthy. We thus obtain the following implication: if a group is collectively blameworthy, then it does not fulfil a subjective collective obligation.

Suppose a group $\mathcal{H}$ is collectively blameworthy. According to the previously mentioned implication, this entails that the group failed to fulfil a subjective collective obligation. Under the assumption that ought implies can, this entails that the group collectively knew how to fulfil its subjective collective obligation. Say it would fulfil its subjective collective obligation if and only if it performs a group action that guarantees $\varphi_{\checkmark}$. The 'effectivity' condition of collective know-how shows that the group commonly knew that they would have fulfilled their subjective collective obligation if and only if each member had played her part. Assuming that each member's individual action becomes public, this means that the group can identify the members who failed to play their part and hold them responsible to an appropriate degree.

What kind of reasons might pardon a member from being responsible for not knowingly carrying out her part, which is necessary for the group to fulfil its subjective collective obligation? Suppose agent i fails to knowingly carry out her part in fulfilling a subjective collective obligation, that is, $\neg \mathsf{K}_i\langle i\ \mathsf{stit}\rangle[\mathcal{H}\ \mathsf{stit}]\varphi_{\checkmark}$ holds. Or, equivalently, $\hat{\mathsf{K}}_i[i\ \mathsf{stit}]\neg[\mathcal{H}\ \mathsf{stit}]\varphi_{\checkmark}$ holds, which means that she considers it possible that she herself guarantees that the group fails to fulfil its subjective collective obligation.[78] The 'part know-how' condition of collective know-how entails that she knows how to play her part, that is, $\mathsf{K}_i\Diamond\mathsf{K}_i\langle i\ \mathsf{stit}\rangle[\mathcal{H}\ \mathsf{stit}]\varphi_{\checkmark}$ holds. In other words, she knew that $[i\ \mathsf{stit}]\neg[\mathcal{H}\ \mathsf{stit}]\varphi_{\checkmark}$ might hold, while knowing how to prevent this. Hence, she decided to perform an action which might implicate herself as *causing* the group's failure to fulfil its subjective collective obligation rather than knowingly guaranteeing that she cannot be so implicated.

The 'causing', or $[i\ \mathsf{stit}]$-operator, above is essential.[79] To see this, let us briefly investigate the case where she considers it possible that the subjective collective obligation was not fulfilled regardless of whether she knowingly plays her part,

[78] Not knowingly carrying out her part is consistent with carrying out her part unknowingly. Hence, she might not *know* that she guarantees that the group fails to fulfil its subjective collective obligation.

[79] Causality has been a thorny issue in philosophy. For my current purposes, it is sufficient to assume that when an agent sees to it that φ, then she causes it.

that is, $\hat{\mathsf{K}}_i\neg[\mathcal{H}\ \mathsf{stit}]\varphi_{\checkmark}$ holds regardless.[80] In this case, knowingly playing her part makes no difference to whether she knowingly risks a collective failure, that is, $\hat{\mathsf{K}}_i\neg[\mathcal{H}\ \mathsf{stit}]\varphi_{\checkmark}$ holds regardless of whether she knowingly plays her part. However, when not knowingly playing her part she knowingly risks that she *herself* is a cause of the collective failure. This is formally expressed by $\hat{\mathsf{K}}_i[i\ \mathsf{stit}]\neg[\mathcal{H}\ \mathsf{stit}]\varphi_{\checkmark}$. Or, conversely, if she chooses to knowingly play her part she at least knows that she herself will not be causally responsible for the collective failure, despite the fact that she does not know that collective failure will be avoided.

A typical context in which a group member knows that the subjective collective obligation is not fulfilled, regardless of whether she herself knowingly plays her part, is when she knows that other group members will not comply. Because of the 'effectivity' condition, if she knows that some group member will not do her part, then she knows that the group will fail to fulfil its subjective collective obligation. To illustrate this possibility, consider the version of the prisoners' dilemma depicted by the STIT model in Figure 3.8.[81] Although this model excludes the knowledge of the prisoners, it can be used to illustrate that an agent may know that the group fails to fulfil its subjective collective obligation, for example, because she knows that others are guided by their personal interests, rather than the collective good. The prisoners' dilemma is a standard game-theoretical example that highlights this worry: each prisoner promotes her personal interests (modelled by u_i or u_j), regardless of what the other prisoner does, only if she defects (modelled by K^D_-). But if each defects, then this will lead to a worse outcome than if each had cooperated (modelled by K^C_-). In fact, to jointly promote the collective good (modelled by $u_{\mathcal{G}}$) they both need to cooperate.

[80] It could even be the case that she knows that the subjective collective obligation is not fulfilled regardless of whether she knowingly plays her part, that is, $\mathsf{K}_i\neg[\mathcal{H}\ \mathsf{stit}]\varphi_{\checkmark}$ holds regardless. The opposite, viz. knowing that the subjective collective obligation is fulfilled while not knowingly carrying out her part, is impossible: it holds that $\mathsf{K}_i[\mathcal{H}\ \mathsf{stit}]\varphi_{\checkmark}$ implies $\mathsf{K}_i\langle i\ \mathsf{stit}\rangle[\mathcal{H}\ \mathsf{stit}]\varphi_{\checkmark}$.

[81] The STIT model relates to what I called a cooperation game in § 2.5. See § 1.1 for an introduction to the theory of games and notational conventions.

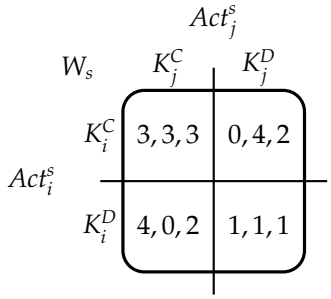


Figure 3.8: *A STIT model of the prisoners' dilemma. The only difference with the standard prisoners' dilemma is that each action profile is assigned a triple of utilities: one for agent i, one for agent j, and one for the group $\mathcal{G} = \{i, j\}$, respectively, where $u_{\mathcal{G}}(w)$ is the average of $u_i(w)$ and $u_j(w)$.*

Why would a member decide against playing her part in jointly promoting the collective good? On the one hand, she may not be motivated to devalue her own personal interests for the collective good. In such cases, there is no individualistic reason for her to knowingly play her part. The individualistic motives of the members of a group may therefore undermine its collective performance.

On the other hand, the member would know that the subjective collective obligation will not be fulfilled if she knows that the others will not impair their personal interests. She would then know that she will not be the sole cause of the collective failure. Having partners in crime might alleviate her individual blameworthiness. However, Matthew Braham and Martin van Hees write:

> a person is an apposite target of censure and sanction for some state of affairs only if ... it was within her power to reasonably choose not to be an "author" (or "co-author" in the presence of multiple causal conditions) of that state of affairs. (Braham and van Hees, 2012, p. 607)[82]

[82]The details of our discussions differ substantially. Their framework includes subjective probabilities, relies on the NESS condition (which stands for 'Necessary Element of a Sufficient Set') to

If we take it that authorship of agent i for χ is implied by $[i \ \mathsf{stit}]\chi$, then it intuitively seems that knowingly causing a collective failure makes one an appropriate target of blame or sanction for the collective failure. A group member may therefore value the fact that she knows that she is not an *author*, or co-author, of the group's failure to fulfil its subjective collective obligation, regardless of whether she knows that the group will fail to do so. Hence, if group members collectively know how to fulfil its subjective collective obligation and if they are motivated in this way, then it is guaranteed that the group will fulfil its subjective collective obligation. In other words, it is impossible for a group to fail to fulfil its subjective collective obligation if each group member values the fact that she knows that she is not an author of the group's failure.

In sum, a group's blameworthiness *logically implies* that at least one member knowingly risked that she herself causes the group's failure to fulfil its corresponding subjective collective obligation. In fact, it logically implies that every member who failed to knowingly play her part is so implicated. From the collective perspective, assuming that the members' individual actions become public, the group can identify the members who failed to play their part and hold them responsible to an appropriate degree. From the individualistic perspective, this means that at least one member knowingly risked that she is an author, or coauthor, of the group's failure to fulfil its subjective collective obligation. I highlighted that individualistic reasons for this divergence may be epistemic, such as when she knows that some other members will also not do their part, or motivational, such as when she does not want to devalue her individual interests for the collective good.[83]

assess causal connections, and includes eligible options. Mine, in contrast, is non-probabilistic, relies on control (or α-effectivity) to assess causal connections, and does not include the eligibility of the available options.

[83] It remains to be studied whether these reasons *justify* the member's divergence and thereby alleviate her individual blameworthiness.

It is important and interesting to develop a full-fledged theory of subjective obligations – individual and collective – and investigate the relation between subjective individual obligations and subjective collective obligations further. For now, I have to settle for these preliminary observations and leave such a study for another occasion.

This page intentionally contains only this sentence.

Appendix C

Collective Know-how

C.1 Epistemic STIT Theory: Proofs

Theorem 3.1 (Completeness Epistemic STIT)**.** *The following axiom schemas, in combination with a standard axiomatization for propositional logic, and the standard rules* (*such as necessitation*) *for the normal modal operators, provide a complete Hilbert system for the validities on epistemic STIT models:*

(S5 Historical Necessity)	*S5 for* $\Box$
(S5 Agency)	*for each group* $\mathcal{H}$*: S5 for* $[\mathcal{H}\ \mathsf{stit}]$
(Agent Monotonicity)	*for all groups* $\mathcal{F}$ *and* $\mathcal{G}$ *satisfying* $\mathcal{F} \subseteq \mathcal{G}$*:* $[\mathcal{F}\ \mathsf{stit}]\varphi \rightarrow [\mathcal{G}\ \mathsf{stit}]\varphi$
(Independence of Agency)	*for all groups* $\mathcal{F}$ *and* $\mathcal{G}$ *satisfying* $\mathcal{F} \cap \mathcal{G} = \emptyset$*:* $\Diamond[\mathcal{F}\ \mathsf{stit}]\varphi \wedge \Diamond[\mathcal{G}\ \mathsf{stit}]\psi \rightarrow \Diamond([\mathcal{F}\ \mathsf{stit}]\varphi \wedge [\mathcal{G}\ \mathsf{stit}]\psi)$
(S5 Knowledge)	*for each* $i \in Ags$*: S5 for* K_i
(Public Knowledge)	*for each group* $\mathcal{H}$*:* $\mathsf{C}_{\mathcal{H}}\varphi \rightarrow (\varphi \wedge \mathsf{E}_{\mathcal{H}}\mathsf{C}_{\mathcal{H}}\varphi)$
(Induction)	*for each group* $\mathcal{H}$*:* $\varphi \wedge \mathsf{C}_{\mathcal{H}}(\varphi \rightarrow \mathsf{E}_{\mathcal{H}}\varphi) \rightarrow \mathsf{C}_{\mathcal{H}}\varphi$

Proof. The epistemic STIT logic is a so-called fusion of epistemic logic and non-epistemic STIT theory. The complete logical system for epistemic STIT theory is therefore given by the simple combination of the logical systems for epistemic logic and (non-epistemic) STIT logic.

It is well-established that the non-epistemic fragment is complete with respect to STIT models, for instance, by using Sahlqvist correspondence (see the standard textbook treatment of Blackburn et al. (2001)). The completeness of the epistemic logic is standard (see, e.g. Meyer and van der Hoek, 1995). (Broersen (2011a, Theorem 2.1) proves a similar result for XSTIT, the difference is that his logic concerns XSTIT and excludes common knowledge.) □

C.2 Individual Practical Knowledge: Proofs

Proposition 3.1.

(OAC) *The* own-action condition*, schematically expressed by* $\mathsf{K}_i\varphi \rightarrow \mathsf{K}_i[i \text{ stit}]\varphi$*, corresponds to the condition:*

For all dynamic states $\langle s,h\rangle$*,* $\langle s',h_1\rangle$*, and* $\langle s',h_2\rangle$*, if* $\langle s,h\rangle \sim_i \langle s',h_1\rangle$ *and* $h_2 \in Act_i^{s'}(h_1)$ *then* $\langle s,h\rangle \sim_i \langle s',h_2\rangle$*.*

(Unif-H) *The* uniformity of historical possibility property*, schematically expressed by* $\Diamond\mathsf{K}_i\varphi \rightarrow \mathsf{K}_i\Diamond\varphi$*, corresponds to the confluency condition:*

For all dynamic states $\langle s,h_1\rangle$*,* $\langle s,h_2\rangle$*, and* $\langle s',h_1'\rangle$*, if* $\langle s,h_1\rangle \sim_i \langle s',h_1'\rangle$ *then there is a history* h_2' *such that* $\langle s,h_2\rangle \sim_i \langle s',h_2'\rangle$*.*

Proof. These correspondences can be checked using the algorithm SQEMA (Conradie et al., 2006). □

Theorem 3.2 (Impossibility Results)**.** *Implicit individual know-how and explicit individual know-how cannot be characterized in the current logic.*

That is, given an agent i and a formula φ, there is no *formula ψ such that for every model $\mathcal{M}$ and every dynamic state $\langle s, h\rangle$ it holds that, relative to $\langle s, h\rangle$, agent i implicitly knows how to φ if and only if $\mathcal{M}, \langle s, h\rangle \vDash \psi$. And something similar applies to explicit know-how.*

Proof. Implicit individual know-how would be characterizable in $\mathfrak{L}_{ESTIT}$ if and only if, given an agent $i \in Ags$ and a formula φ, there is a formula ψ available such that for any model $\mathcal{M}$ and any dynamic state $\langle s, h\rangle$ it holds that $\mathcal{M}, \langle s, h\rangle \vDash \psi$ iff i implicitly knows how to φ.

We argue by contradiction. Suppose ψ characterizes that agent i implicitly knows how to p. Let $P' \subset P$ denote the set of propositional variables occurring in either ψ or p, and let $q \notin P'$. Consider a model $\mathcal{M}$ and a dynamic state $\langle s, h\rangle$ in which agent i has two options, that is, Act_i^s contains two acts. Let the epistemic indistinguishability relation $\sim_i$ be given by $w \sim_i v$ if and only if $w \in Act_i^s$. Let the valuation function V assign to each propositional variable the set of all dynamic states in the model. It is now easy to see that i implicitly knows how to p, because for every formula χ we have $\mathcal{M} \vDash \Box(\varphi \leftrightarrow \chi)$ or $\mathcal{M} \vDash \Box(\varphi \leftrightarrow \neg\chi)$. The assumption implies that $\mathcal{M}, \vDash \psi$ holds. See Figure C.1 for the initial model in which agent i implicitly knows how to φ.

(a) *The initial model $\mathcal{M}$.* **(b)** *The constructed model $\mathcal{M}'$.*

Figure C.1

We can construct a countermodel $\mathcal{M}'$ from $\mathcal{M}$ by adjusting the propositional valuation to V' in the following way: one choice guarantees q, the other guarantees $\neg q$, and both guarantee any $p' \in P'$ (see the constructed model in Figure C.1). Note

that V' only differs from V on q. It is easy to see that in $\mathcal{M}'$, i does not implicitly know how to φ, because, in $\mathcal{M}'$, q is a proper refinement of p for agent i. However, because we did not alter the truth values of the propositions occurring in ψ, it is the case that ψ also holds in $\mathcal{M}'$. Contradiction.

In a similar way, one can prove that explicit knowledge is also not characterizable. □

Theorem 3.3 (Characterization of Individual Know-how)**.** *Individual know-how is characterized by*

$$\mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi.$$

That is, for every model $\mathcal{M}$, every dynamic state $\langle s,h\rangle$, every individual agent $i \in Ags$, and every formula φ it holds that, relative to $\langle s,h\rangle$, agent i knows how to φ if and only if $\mathcal{M},\langle s,h\rangle \vDash \mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi$.

Proof. (**If.**) Follows immediately by letting $\psi := \varphi$ in Definition 3.5.

(**Only if.**) It is immediate for implicit know-how. For explicit know-how it follows from two facts. First, $\Box$ is monotonic: $\vDash \Box(p \to q) \to (\Box p \to \Box q)$, which entails $\vDash \Box(p \to q) \to (\Diamond p \to \Diamond q)$. Replacing p with $\mathsf{K}_i[i\ \mathsf{stit}]\psi$ and q with $\mathsf{K}_i[i\ \mathsf{stit}]\varphi$ yields

$$\vDash \Box(\mathsf{K}_i[i\ \mathsf{stit}]\psi \to \mathsf{K}_i[i\ \mathsf{stit}]\varphi) \to (\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\psi \to \Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi).$$

Second, by K_i-necessitation, and two applications of the K-axiom for K_i it holds that:

$$\vDash \mathsf{K}_i\Box(\mathsf{K}_i[i\ \mathsf{stit}]\psi \to \mathsf{K}_i[i\ \mathsf{stit}]\varphi) \to (\mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\psi \to \mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi).$$

By assumption, both the antecedent of the principal implication and the antecedent of the subordinate implication obtain. Hence, $\mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi$ holds, as desired. □

C.3 Collective Know-how: Proofs

Theorem 3.4. *Let $\mathcal{M}$ be an epistemic STIT model, and let $\langle s, h\rangle$ be a dynamic state. Assume that $\mathcal{M}$ satisfies (Unif-H). Suppose $\mathcal{H}$ collectively knows how to φ, as witnessed by $(\varphi_i)_{i\in\mathcal{H}}$. Then*

$$\mathcal{M}, \langle s, h\rangle \vDash \mathsf{C}_{\mathcal{H}}\Box \bigwedge_{i\in\mathcal{H}}([i\ \mathsf{stit}]\varphi_i \leftrightarrow \langle i\ \mathsf{stit}\rangle[\mathcal{H}\ \mathsf{stit}]\varphi), \qquad (*)$$
$$\mathcal{M}, \langle s, h\rangle \vDash \mathsf{C}_{\mathcal{H}}\Box \bigwedge_{i\in\mathcal{H}}(\mathsf{K}_i[i\ \mathsf{stit}]\varphi_i \leftrightarrow \mathsf{K}_i\langle i\ \mathsf{stit}\rangle[\mathcal{H}\ \mathsf{stit}]\varphi). \qquad (**)$$

Proof. Assume all the stated assumptions. To clarify, the crucial assumptions are (1) $\mathsf{C}_{\mathcal{H}}\Box(\bigwedge_{i\in\mathcal{H}}[i\ \mathsf{stit}]\varphi_i \leftrightarrow [\mathcal{H}\ \mathsf{stit}]\varphi)$ and (2) $\mathsf{C}_{\mathcal{H}} \bigwedge_{i\in\mathcal{H}} \mathsf{K}_i\Diamond\mathsf{K}_i[i\ \mathsf{stit}]\varphi_i$. It is important to note that (2) entails (2′) $\mathsf{C}_{\mathcal{H}} \bigwedge_{i\in\mathcal{H}} \Diamond[i\ \mathsf{stit}]\varphi_i$.

First, from the right-to-left implication in (1) and the fact that $\vDash \langle i\ \mathsf{stit}\rangle[i\ \mathsf{stit}]\varphi_i \to [i\ \mathsf{stit}]\varphi_i$, it immediately follows that $\mathsf{C}_{\mathcal{H}}\Box \bigwedge_{i\in\mathcal{H}}(\langle i\ \mathsf{stit}\rangle[\mathcal{H}\ \mathsf{stit}]\varphi \to [i\ \mathsf{stit}]\varphi_i)$. Conversely, because of the left-to-right implication in (1), $[j\ \mathsf{stit}]\varphi_j$ would be compatible with $[\mathcal{H}\ \mathsf{stit}]\varphi$ if all other members were able to play their part. Due to (2′), all other members are able to play their part. Hence $\mathsf{C}_{\mathcal{H}}\Box \bigwedge_{i\in\mathcal{H}}([i\ \mathsf{stit}]\varphi_i \to \langle i\ \mathsf{stit}\rangle[\mathcal{H}\ \mathsf{stit}]\varphi)$.

Second, $(**)$ follows from $(*)$ and two facts. First, the fact that $\vDash \mathsf{C}_{\mathcal{H}}\chi \to \mathsf{C}_{\mathcal{H}}\mathsf{K}_i\chi$, for each $i \in \mathcal{H}$. Second, the fact that (Unif-H) entails that $\mathcal{M} \vDash \mathsf{K}_i\Box\chi \to \Box\mathsf{K}_i\chi$. □

Observation 3.3 (Interchangeability & Effectivity)**.** *Let S be a game model, let $\mathcal{G}$ be a group of agents, and let φ be a formula. The following are equivalent:*

1. $S \vDash \Box([\mathcal{G}\ \mathsf{stit}]\varphi \leftrightarrow \bigwedge_{i\in\mathcal{G}}\langle i\ \mathsf{stit}\rangle[\mathcal{G}\ \mathsf{stit}]\varphi)$*;*
2. $P_{\mathcal{G}} := \{a_{\mathcal{G}} \in A_{\mathcal{G}} \mid S, a \vDash [\mathcal{G}\ \mathsf{stit}]\varphi\}$ *is interchangeable.*

Proof. (1. $\Rightarrow$ 2.) Assume 1. Let $b, c \in A$ be such that $b_{\mathcal{G}}, c_{\mathcal{G}} \in P_{\mathcal{G}}$, and let $i \in \mathcal{G}$. To prove 2., we take an arbitrary $d \in A$ such that $d_{\mathcal{G}} = (b_{\mathcal{G}-i}, c_i)$, and prove that $d_{\mathcal{G}} \in P_{\mathcal{G}}$. For any $j \in \mathcal{G} - i$ it holds that $S, d \vDash \langle j\ \mathsf{stit}\rangle[\mathcal{G}\ \mathsf{stit}]\varphi$, because $S, b \vDash [\mathcal{G}\ \mathsf{stit}]\varphi$

and $d_j = b_j$. Hence, $S, d \vDash \bigwedge_{j \in \mathcal{G}-i} \langle j \text{ stit} \rangle [\mathcal{G} \text{ stit}] \varphi$. In a similar way, we can prove that $S, d \vDash \langle i \text{ stit} \rangle [\mathcal{G} \text{ stit}] \varphi$ (note the 'i' here) . By the right-to-left implication in 1. it follows that $S, d \vDash [\mathcal{G} \text{ stit}] \varphi$. Hence $d_{\mathcal{G}} \in P_{\mathcal{G}}$.

(2. $\Rightarrow$ 1.) Assume 2. Take any $a \in A$. We need to show that $S, a \vDash [\mathcal{G} \text{ stit}] \varphi \leftrightarrow \bigwedge_{i \in \mathcal{G}} \langle i \text{ stit} \rangle [\mathcal{G} \text{ stit}] \varphi$. The left-to-right implication follows immediately from the fact that $[i \text{ stit}]$ is reflexive, i.e. satisfies $\vDash \chi \rightarrow \langle i \text{ stit} \rangle \chi$. To establish the right-to-left implication, assume $S, a \vDash \bigwedge_{i \in \mathcal{G}} \langle i \text{ stit} \rangle [\mathcal{G} \text{ stit}] \varphi$. That is, for every $i \in \mathcal{G}$ there is a $b^i_{\mathcal{G}} \in P_{\mathcal{G}}$ such that $b^i_i = a_i$. Using the fact that $P_{\mathcal{G}}$ is interchangeable, one can inductively show that for any subgroup $\mathcal{F} \subseteq \mathcal{G}$ there is a $b^{\mathcal{F}}_{\mathcal{G}} \in P_{\mathcal{G}}$ such that $b^{\mathcal{F}}_{\mathcal{F}} = a_{\mathcal{F}}$. Hence, $a_{\mathcal{G}} = b^{\mathcal{G}}_{\mathcal{G}} \in P_{\mathcal{G}}$ and therefore there is an $a' \in A$ such that $a'_{\mathcal{G}} = a_{\mathcal{G}}$ and $S, a' \vDash [\mathcal{G} \text{ stit}] \varphi$. Since $\vDash [\mathcal{G} \text{ stit}] \varphi \leftrightarrow [\mathcal{G} \text{ stit}][\mathcal{G} \text{ stit}] \varphi$ and $a'_{\mathcal{G}} = a_{\mathcal{G}}$, it holds that $S, a \vDash [\mathcal{G} \text{ stit}] \varphi$, as desired. □

4

Joint Action, Participatory Intentions, and Team Reasoning

> *Call this way of conceiving of action a participatory intention: an intention to do my part of a collective act, where my part is defined as the task I ought to perform if we are to be successful in realizing a shared goal. This conception of oneself as contributing to a collective, as manifested in one's deliberation and action, is what lies at the heart of collective action generally, from simple coordination to complex cooperation.*
>
> Christopher Kutz (2000, p. 81)

4.1 Introduction

When a group of agents has come together to strive for a collective or joint goal, they must act in such a way that their joint action promotes the realization of their collective goal. What kind of personal objective should the members of a group

[†]This chapter is largely based on Duijf (forthcoming a), although many parts have been thorougly revised or expanded (the most significant additions are the figures in § 4.4 and the discussions in § 4.4.4 and § 4.6). It is heavily influenced by my published joint work with Allard Tamminga (Tamminga and Duijf, 2017) and my work with Jan Broersen and John-Jules Meyer (Duijf et al., forthcoming).

adopt to guarantee that they together promote the realization of their collective goal? What is the relation between a group's collective intention and its members' individual intentions?

To address these questions, I combine modal-logical and game-theoretical formalisms to study various types of intentions that a group member could plausibly adopt when the group collectively aims to achieve a collective goal. My enquiry is hence in line with the philosophical literature on collective intentionality, which typically analyses a group's collective intention as an interlocking web of its members' individual intentions. Let us call such individual intentions *we-intentions*, as they relate to the group's collective intention. I characterize and investigate three types of we-intentions: so-called *pro-group*, *team-directed*, and *participatory intentions*.[1] Roughly stated, a member who adopts a pro-group intention can be viewed as adopting the collective goal as her own personal goal; a member who adopts a team-directed intention can be thought of as aiming to perform an individual action that is compatible with a best joint act; and, finally, a member who adopts a participatory intention can be taken to aim at realizing that a best joint action is performed. To clarify that the pro-group intention is different from the other two types of we-intentions, it may be helpful to add that in cases of choice under uncertainty the collective objective may be realized even though the group does not perform a group action that best promotes the realization of its objective, and vice versa. To elucidate the distinction between team-directed intentions and participatory intentions, note that if there are several best group acts available then it could be that an agent performs an individual action that is compatible with a best joint action even though there is an individual action available to her that better promotes the realization of a best joint act.[2] That

[1] Although 'pro-group' and 'team-directed' are hardly distinguishable in plain English, it will become clear in § 4.2 that my terminology tracks the distinction between so-called 'pro-group I-mode reasoning' and 'team-directed reasoning'.

[2] For example, an individual action that guarantees that a best group action is performed promotes the realization of a best joint act better than an individual action that is only compatible with a best joint act.

is, the action recommendations yielded by participatory intentions refine those yielded by team-directed intentions. Although the details of these three types of we-intentions are important, they need not worry us at present. One of this chapter's vital contributions is to systematically explain the intricate differences between these we-intentions using logical machinery (§ 4.4).

One may think that the group will jointly select an optimal group action regardless of the type of we-intention that its members adopt. I show that this is, however, not the case. The adoption of, for example, pro-group intentions may produce different group acts from the adoption of participatory intentions. It is therefore important to investigate for each type of we-intention the class of games for which it guarantees successful cooperation. My central results show that participatory intentions *surpass* both team-directed and pro-group intentions in guaranteeing successful cooperation (§ 4.5, visualized in Figure 4.1). That is, in *any* scenario if either team-directed or pro-group intentions guarantee that a best joint action is performed, then so do participatory intentions. So when a group of agents strives to achieve some joint goal, its members should adopt participatory intentions, rather than team-directed or pro-group intentions.

I mentioned that the philosophical literature typically conceptualizes a group's collective intention as an interlocking web of its members' individual intentions. A strand of literature on *team reasoning* (Sugden, 2000; Bacharach, 2006; Gold and Sugden, 2007, see § 2.5.1) opposes this trend and claims that approaches that focus on we-intentions miss what is inherently cooperative about joint actions because they fail to appeal to the reasoning process leading up to the formation of these we-intentions. To distinguish a random set of individual actions from cooperation, team-reasoning theorists therefore contrast individualistic reasoning and team reasoning. Raimo Tuomela's philosophical theory of sociality (Tuomela, 2000a, 2005, 2006, 2007, see § 4.2.2) relies on a similar distinction between *I-mode reasoning* and *we-mode reasoning*. My first result establishes that an important part

of the we-mode reasoning akin to team reasoning can be *reduced* to individualistic reasoning with a team-directed intention. This is important because it opposes Bacharach (1999) and Hakli, Miller, and Tuomela (2010).[3] Moreover, in cases where there are several best group actions available, team reasoning can be *refined* by taking the resulting intentions to be participatory intentions rather than team-directed intentions.

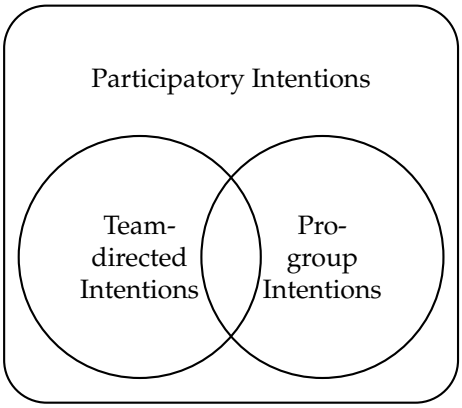


Figure 4.1: *Cooperating successfully: each area depicts the class of games for which the respective intention type guarantees successful cooperation.*

I will illustrate my formalism with ideas and concepts from the philosophical and economics literature on intentionality, team reasoning, cooperation, and joint action. This serves three purposes. First, the concepts and ideas from philosophy and economics help to clarify what the formalism is meant to model and thereby gives a partial conceptual justification for my study of various we-intention types. Second, my study aims to bridge the team-reasoning account of cooperation and philosophical analyses of collective intentionality and thereby provides a

[3]Hakli et al. (2010, Thesis (5) on p. 307) write: "We-mode reasoning is not reducible to pro-group I-mode reasoning, i.e. it is not definable by or functionally constructable from I-mode reasoning." These opposing results derive from the fact that Bacharach (1999) and Hakli et al. (2010) assume that I-mode reasoning is connected to equilibrium-based reasoning, whereas I assume that I-mode reasoning is connected to dominance reasoning.

fruitful framework for future interaction and cross-fertilization between these fields. Third, I argue that the formalism captures important aspects of concepts and ideas developed in the literature and therefore my discussion and formal results are of significance for the debates on team reasoning, cooperation, and we-intentions. To highlight this impact, I discuss the implications for philosophical analyses of collective intentionality (§ 4.2.2), and briefly elaborate on implications for game-theoretical frameworks of cooperation (§ 4.4.4). In the final section I return the overall theme of this thesis and discuss the relation between individual and collective blameworthiness from the perspective of my formalism.

The chapter is set out as follows. In § 4.2, I briefly discuss team reasoning and prominent philosophical accounts of collective intentionality to highlight and clarify the aim of this chapter. In § 4.3, I then extend traditional game forms with intentions and introduce a modal logic of agency and intentionality, which is interpreted on these game forms. In § 4.4, the modal-logical machinery is used to accurately characterize the three types of we-intentions – pro-group, team-directed, and participatory intentions. In addition, I show that team-directed intentions yield the same action recommendations as team reasoning. I discuss the implications of my work for the study of fairness and cooperation in game theory. In § 4.5, I investigate the classes of games for which the three types of we-intentions guarantee successful cooperation. There are two central results: (1) Pro-group intentions and team-directed intentions are on a par: in some scenarios team-directed intentions succeed in guaranteeing the realization of a best group action whereas pro-group intentions fail to do so, and vice versa. The class of games for which team-directed intentions guarantee successful cooperation is hence not a proper superset of the corresponding class for pro-group intentions, nor the other way around. (2) Participatory intentions surpass both pro-group intentions and team-directed intentions: in any scenario, if either team-directed or pro-group intentions guarantee that a best group action is performed, then

so do participatory intentions. Or, equivalently, the class of games for which participatory intentions guarantee successful cooperation is a superset of both the corresponding class for pro-group intentions and the corresponding class for team-directed intentions. The theory of participatory intentions is therefore the prevalent account of cooperation. In the concluding section, I briefly discuss the relation between collective and individual responsibility from the perspective of my theory of participatory intentions.

4.2 Team Reasoning and Collective Intentionality

The philosophical debate on collective intentionality has recently been enriched by the economics literature on team reasoning. On the one hand, philosophical accounts typically rely on members to form a *personal intention* of sorts.[4] On the other hand, the economic literature on team reasoning claims to yield a better understanding of collective intentionality (Gold and Sugden, 2007; Hakli et al., 2010). Instead of relying on members' personal intentions, the latter approach appeals to the *reasoning process* leading up to the formation of the intention and argues that this focus is needed to grasp collective intentionality. In short, philosophers focus on the intentions of the members, while economists focus on their reasoning process. Although these accounts seem to differ fundamentally, one of the goals of this chapter is to show that team reasoning yields the same action recommendations as one of the we-intentions, viz. team-directed intentions (Result 1). I will briefly present these accounts to clarify the scope and limits of my study with regard to team reasoning and to highlight the implications for the philosophical debate on collective intentionality.

[4]Philosophical accounts of collective intentionality build on participatory intentions (Kutz, 2000), contributory intentions (discussed and rejected by Gilbert (2009)), intentions that we *J* (Bratman, 2014), and we-intentions (Tuomela, 2000a, 2005; Searle, 1990).

4.2.1 Team Reasoning

The idea of team reasoning originates from ethical theories of utilitarianism. Hodgson (1967) used it to demonstrate that rule and act utilitarianism rely on different modes of reasoning. Regan (1980) later expanded on this argument in his theory of 'cooperative utilitarianism'. In the nineties Sugden (1991, 1993) has fruitfully introduced team reasoning to the field of game theory. The core idea of team reasoning is that a member of a group asks herself 'What should *we* do?' rather than 'What should *I* do?'. Team reasoning hence relies on a we-perspective. This means that a team reasoner first considers the group actions available to the group, assesses these group actions in terms of their consequences, finds the group action that best furthers their common or collective interests, and then chooses her component of that group action.[5]

To explain team reasoning in more detail and to contrast it with traditional rational choice theory, let us consider the Hi-Lo game depicted in Figure 4.2. Team-reasoning theorists claim that traditional rational choice theory does not adequately address the Hi-Lo game. It seems that (*high*, *high*) is the only rational solution, but the players have no reason for preferring one action over the other if they are guided by traditional rational choice theory. Let us see why. It is evident that player 1 should choose *high* if she expects player 2 to choose *high* too. Likewise, she should choose *low* if she expects her opponent to choose *low*. In particular, according to expected utility theory, a player should choose *high* if and only if she judges the probability that her opponent chooses *high* to be greater than ⅓.[6] This means that what is rational for player 1 depends on what player 2 can be expected to choose. At best, this gives conditional recommendations.

[5]Similar ideas have been proposed by Anderson (2001); Hurley (1989); and Gilbert (1989).

[6]The Principle of Indifference (often attributed to Keynes, 1921, Ch. 4) may be invoked to argue that player 1 should assign equal probability to player 2 choosing *high* and player 2 choosing *low*. As desired, this would then entail that that she should choose *high*. However, I seriously doubt that the application of the Principle of Indifference can be justified in the Hi-Lo game as a principle of rationality. Simply stated, I think its application is only compatible with theories of *bounded* rationality (see also Bacharach, 2006, § 1.6).

At worst, it gives her an indeterminate unconditional recommendation. Similar objections apply to dominance reasoning. In such cases game theorists typically invoke the assumption of common knowledge of rationality.

What can player 1 expect player 2 to do under the assumption of common knowledge of rationality? One may think that the Nash equilibrium concept captures this assumption.[7] An action profile is a Nash equilibrium if and only if no player can be better off by unilaterally changing her component action (see Definition 1.3). It is a state of equilibrium because no one has an incentive to deviate from performing her component individual action, given that everyone else performs their respective part. The Hi-Lo game, however, contains two (pure) Nash equilibria: (*high*, *high*) and (*low*, *low*).[8] As a response to this multiplicity, one may want to refine the Nash equilibria by appealing to the Pareto dominance of (*high*, *high*) over (*low*, *low*) in order to select the Nash equilibrium (*high*, *high*) as the only rational solution.[9] There are two problems with such a solution. First, what is the status of the Pareto principle in standard rational choice theory? Hollis and Sugden (1993, p. 13) argue that the Pareto principle "is a principle of rationality only to players who conceive of themselves as a team, but not for players who do not". It is therefore plausible that the rationalization of the Pareto principle requires a departure from the central assumption in rational choice theory that agency is only invested in individuals. Second, in any case, such a (Pareto-dominant) Nash equilibrium only captures a possible status-quo: *if* everyone expected the others to play their part in the Nash equilibrium, *then* they would have a reason to do the same. It hence gives only a conditional recommendation

[7] In fact, the initial question prompted the field of epistemic game theory (see the textbook by Perea, 2012). Among others, such developments have led to a concept of rationalizability (Bernheim, 1984; Pearce, 1984), which is more general than the Nash equilibrium concept. In particular, epistemic game theorists have disproved the claim that common knowledge of rationality entails Nash equilibrium play.

[8] I will restrict my discussion to pure strategies. This restriction is harmless as mixed strategies do not succeed in addressing the present worries.

[9] Harsanyi and Selten (1988) argue for Pareto dominance as a principle of equilibrium selection.

		Player 2: High	Player 2: Low
Player 1	High	2 2	0 0
	Low	0 0	1 1

Figure 4.2: *The Hi-Lo game.*

and triggers an infinite regress of reasons.[10] This inadequate response to the Hi-Lo game by traditional rational choice theory stands to be corrected, which is what team reasoning (as studied by Bacharach, Sugden, and Gold) has been designed for.

Bacharach (2006, Ch. 1) and Sugden (2000, §§ 2, 3, 7 and 8) argue that traditional rational choice theory needs to be augmented with a collectivistic reasoning method to successfully address the Hi-Lo game. Team-reasoning theorists appeal to the *reasoning process* by which an individual agent reasons about what to do. An individual agent engaged in team reasoning "works out the best feasible combination of actions for all the members of her team, then does her part in it" (Bacharach, 2006, p. 121). In the Hi-Lo game, this reasoning goes as follows: the row player first identifies (*high*, *high*) as the best combination of individual actions that they can perform and then decides to perform her part in that combination, i.e. *high*. Similar reasoning prescribes *high* for the column player. Team reasoning therefore entails that *high* is the only rational option, and selects (*high*, *high*) as the only rational outcome. Problem solved.[11]

[10] Hollis and Sugden (1993), Sugden (2000, pp. 179–182) and Bacharach (2006, pp. 35–68) provide more elaborate treatments of these objections to traditional rational choice theory.

[11] Bardsley et al. (2010) present ample experimental evidence that people generally agree that *high* is the right thing to do.

To help clarify the scope and limits of my study, let me briefly discuss three questions that the literature on team reasoning needs to address. First, under which conditions should or will an individual agent team reason rather than reason individualistically? Michael Bacharach's (2006) theory of team reasoning is constructed within his variable frame theory. On his view, "among the many different dimensions of the frame of a decision-maker is the 'unit of agency' dimension: the framing agent may think of herself as an individual doer or as part of some collective doer" (p. 137). Group identification is one of the key factors for determining whether an agent adopts the 'we-frame', which, according to Bacharach, may be prompted by psychological factors.[12] Robert Sugden (2003), in contrast, thinks that endorsing team reasoning may depend on an assurance that others will likewise endorse it. He writes: "team reasoning does not generate reasons for choice unless each member of a team has reason to believe that there is common reason to believe that each member of the team endorses and acts on team reasoning. This is a condition of assurance" (pp. 176–177). The details of this condition are spelled out in terms of a Lewisian analysis of reason to believe (for a useful summary, see Gold and Sugden, 2007, § III.3). Although this does not settle the issue, my brief discussion shows that there are diverging views on what triggers team reasoning. For my current purposes, it suffices to assume that team reasoning is prompted in the context of collective intentional action.[13]

Second, what is the group's collective objective? Team reasoning is generally taken to presuppose a group preference (see Bacharach, 1999; Sugden, 2000). There are only a handful of proposals in the literature that specify what this group preference generally involves. Sugden (2010, 2011, 2015) seems to rely on

[12]Because it may be uncertain which individuals adopt a we-frame, it may be helpful to add that Bacharach (1999, § 2) presents a model of 'unreliable team interactions', which is meant to capture this kind of uncertainty.

[13]Sugden (2003, p. 168) writes: "By 'team reasoning, narrowly defined' I mean a mode of reasoning, followed by one individual, which prescribes that he should perform his part of whichever profile is best for the team. This mode of reasoning may be embedded in a larger logic which specifies the conditions under which team reasoning, narrowly defined, should be used." My analysis is thus limited to 'team reasoning, narrowly defined'; I do not study the larger logic.

a notion of mutual advantage or benefit.[14] Bacharach (2006, p. 59), on the other hand, hypothesizes that a team reasoner "ranks all act-profiles, using a Paretian criterion". There has been some discussion on the relation between the group preference and the individual preferences. For example, Gold (2012, p. 195) writes that Bacharach "allowed in principle that the group objective might be welfare decreasing for some members" and according to Sugden (2000, p. 176) "the preferences of a team are not necessarily reducible to, or capable of being constructed out of, the preferences that govern the choices that the members of the team make as individuals". In the context of collective intentional action, it seems plausible that the group's collective objective is given. Therefore I will assume that the collective objective is specified. My study is hence independent of any particular account of group preferences.

It is helpful to add that team reasoning, interpreted strictly, does not operate on the Hi-Lo game, as represented in Figure 4.2, because the game only shows the individual preferences. Because team reasoning relies on group preferences, we need to make some assumption regarding these group preferences. It is nonetheless uncontroversial that the group prefers (*high*, *high*) over (*low*, *low*). Team reasoning thus yields a convincing argument for choosing *high* in the Hi-Lo game.[15]

There is a subtle difference between team reasoning and team-directed reasoning. To clarify team reasoning, Sugden (2000, p. 195) writes, for a given group $\mathcal{G}$: "Suppose the following two conditions are satisfied. First, each individual $i \in \mathcal{G}$ engages in team-directed reasoning with respect to $\mathcal{G}$ and [the team-directed preferences]. Second, each individual $i \in \mathcal{G}$ has full team confidence with respect to

[14] Karpus and Radzvilas (2018) develop a detailed formal account of mutual advantage.

[15] In recent years, Sugden (2015, p.156) seems to embrace a theory of team reasoning that relies on notions of *mutual advantage* and *conventions*. For example, he writes: "If individuals are to cooperate effectively, they need to be ready to play their parts in mutually beneficial practices that seem to them to be – and perhaps really are – less than ideal." His new theory of team reasoning is, for instance, consistent with (*low*, *low*)-play in the Hi-Lo game. Nonetheless, I will stick to the original idea that team reasoning is designed to explain why *high* is uniquely rational.

$\mathcal{G}$ and [the team-directed preferences]. Then ... the team engages in *team reasoning*." Moreover, in these circumstances, the team-directed preferences represent the group preferences. To explain the difference between team-directed and team reasoning, consider the following description by Sugden (2000, p. 195): "Team-directed reasoning is something that one individual can engage in, independently of any others. Similarly, team-directed preferences are preferences that can be held by any individual, independently of any others. In contrast, team reasoning, team preferences and team agency are properties of *a set of* individuals, and require a network of common beliefs." Since I will not study the network of common beliefs required for full team confidence, and because I study individual agents' independent reasoning, my study can be viewed as focusing on team-directed reasoning. However, because my study is cast against the background of collective intentional action, it seems plausible that the collective objective is given and that the condition of full team confidence is met. Hence, my discussions and results also pertain to team reasoning.

Third, how can the theory of team reasoning be extended to apply to a greater variety of problems – not just the Hi-Lo game?[16] This is exactly my concern in this chapter. By studying the logical form of the we-intentions that may result from team reasoning, I provide a framework for investigating a wide range of problems. Most importantly, my study includes cases where there is no unique best group action. That is, it includes scenarios in which the group preference does not determine a unique best combination of individual actions. I show that team reasoning may be improved by taking the resulting we-intentions to be participatory intentions (§ 4.5).

[16]Compare Bacharach (2006, p. 58 – amended notation): "There are three requirements for a good theory of why people play *high* in Hi-Lo: (i) that it imply observed behaviour, that is, the almost universal choice of *high* in normal circumstances; (ii) that it do so intelligibly to us, which (to the extent that *high* intuitively and stably seems to us the only rational thing to do) involves displaying *high* as uniquely rational – that is, giving principles of rationality which are themselves persuasive, and showing they dictate doing *high*; and (iii) that it be part of a unified theory of a wide range of problems, not just Hi-Lo – for example, all problems of cooperation." My study mainly addresses the third requirement.

4.2.2 Collective Intentionality and We-intentions

To introduce philosophical analyses of collective intentionality it is best to start with one of the most influential accounts. The conditions for a shared intentional activity to *J* are summarized by Michael Bratman as follows:

> (i) we each intend that we *J*.
>
> (ii) we each intend the following: that we *J* by way of the intentions of each that we *J* (and that the route from these intentions to our joint activity satisfies the connection condition).
>
> (iii) we each intend the following: that we *J* by way of meshing subplans of each of our intentions in favour of our *J*-ing.
>
> (vii) there is common knowledge among the participants of the conditions cited in this construction. (p. 60 Bratman, 2014, – amended notation)[17]

Many contemporary theorists disagree with some part of Bratman's analysis, but they largely remain faithful to its conceptual apparatus. Most importantly, a collective intention is typically viewed as an interlocking web of individual intentions.[18] There is much debate on how to view these individual intentions, which has prompted two opposing theories: either suggesting that these individual intentions are held in a particular *mode*, or suggesting that they have a particular *content*.[19] More specifically, mode-based accounts hold that the individual intentions underlying a collective intention are fundamentally different

[17] Unfortunately I cannot do full justice to his elaborate theory, yet this characterization includes Bratman's most vital building blocks and highlights the gist of his approach.

[18] Compare, for example, Tuomela (2005, p. 330): "it can technically be said that a joint intention consists of the participants we-intentions about the existence of which the participants have mutual belief".

[19] There is a third minority view which analyses collective intention and action by appealing to a plural subject. Gilbert (1990, p. 7), for instance, writes: "When a goal has a plural subject, each of a number of persons (two or more) has, in effect, offered his will to be part of a pool of wills which is dedicated, as one, to that goal."

from standard individual intentions; content-based accounts deny this. Bratman's account is best interpreted as a content-based account since he notes that the difference between standard intentions to act and intentions that we *J* "is not between two fundamentally different attitudes, but between two different kinds of contents of the attitude of intending" (Bratman, 2014, p. 14).[20] The three introduced we-intentions – pro-group, team-directed, and participatory intentions – are best viewed as standard intentions that are distinguished by their content. Hence, my theory of we-intentions is best viewed as a content-based account.[21]

To explain the gist of mode-based accounts of collective intentionality and to discuss the implications of my central results, I focus on Tuomela's philosophical theory of sociality (Tuomela, 2000a, 2005, 2006, 2007), which relies on the distinction between the I-mode and the we-mode: "The we-mode involves functioning as a group member and not as a private person while the I-mode is concerned only with functioning as a private person" (Tuomela, 2006, p. 49). Following Hakli et al. (2010, pp. 315–318), it is useful to divide the we-mode reasoning process up into three stages: the first results in the formation of "a group preference matrix", the second reaches "a joint intention to act", and in the third "the agents select their part-actions".[22] I focus solely on the third stage of the we-mode reasoning process.

[20] Compare Kutz (2000, p. 74): "I will defend an account of collective action in which what makes a set of individual acts a case of jointly intentional action is the content of the intentions with which the individuals act."

[21] My use of the term 'participatory intentions' is inspired by Kutz (2000), although I do not study the relation between our notions. Kutz writes: "On the one hand, collective activity is an ineliminable part of the content of agents' participatory intentions. [. . .] On the other hand, participatory intentions are simply a special class of ordinary intentions, differentiated by their group-oriented content" (pp. 85–86). He thus has in mind a standard intention.

[22] Hakli et al. (2010, p. 318) extensively discuss these three stages and argue that the result of the first stage is common knowledge, but the latter two "can be performed by the individual agents autonomously as generally supposed in non-cooperative game theory". Because the latter two stages of we-mode reasoning do not require a network of beliefs, this means that they can be viewed as team-directed reasoning.

My study is thus cast against the background of collective intentional action. For my purposes, the notions of *pro-group I-mode* and *we-mode* in decision making are essential.

> The pro-group I-mode is concerned with promoting the group's interests. (Hakli et al., 2010, p. 296)

That is, a pro-group I-mode agent transforms her preferences and adopts the group's objectives as if they were her own personal objectives. Then she decides what to do by way of individualistic reasoning. In such a case, we say that she forms a *pro-group intention*, meaning that she intends to further the group's objective.[23] Note that adopting a pro-group intention may require her to take the perspective of the group to determine the group's objective.[24]

In contrast, we-mode reasoning naturally relates to team reasoning:

> We-mode reasoning and Bacharach's team reasoning yield the same action recommendations in game-theoretic choice situations. (Hakli et al., 2010, p. 301)

That is, for my purposes, a we-mode reasoner and a team reasoner are the same. I will show that the results of team reasoning can be explained by team-directed intentions, which are best viewed as standard intentions with a particular content (Result 1). The distinction between pro-group I-mode and we-mode can therefore be characterized in my content-based account as the distinction between pro-

[23] Compare Bacharach's (1999, p. 128 – notation adapted) notion of a "group benefactor": "Let us call a type of player who reasons individualistically but whose payoff function coincides with that of a team $\mathcal{G}$ a *benefactor* of $\mathcal{G}$." Hakli et al. (2010, p. 301) argue that "pro-group I-mode reasoning, in cases in which agents adopt the group preferences, and Bacharach's reasoning as a team benefactor yield the same action recommendations." So, a pro-group I-mode reasoner, a group benefactor, and a pro-group intention adopter yield the same action recommendations.

[24] Hakli et al. (2010, Section 3.2) describe the formation of group preferences by means of we-mode reasoning. It thus seems possible that a pro-group I-mode reasoner partially reasons in the we-mode, though not till the very end. To be more precise, a pro-group I-mode reasoner only follows the first stage of we-mode reasoning.

group intentions and team-directed intentions.[25] This means that an important part of the distinguished we-mode reasoning can be reduced to I-mode reasoning with a team-directed intention. This opposes Hakli et al. (2010, Thesis (5) on p. 307): "We-mode reasoning is not reducible to pro-group I-mode reasoning, i.e. it is not definable by or functionally constructable from I-mode reasoning."[26] Moreover, since team-directed intentions can be improved to participatory intentions, an important part of we-mode reasoning can be improved by taking the resulting intentions to be participatory intentions (§ 4.5).

4.3 Formal Preliminaries

4.3.1 Games and Intentions

A game form can be taken to represent an interdependent decision context involving a finite set N of individual agents. Each individual agent i is assigned a finite set of available actions A_i. The Cartesian product $\times_{i \in N} A_i$ of all individual agents' sets of available actions gives the full set A of action profiles. (See § 1.1 for a more detailed introduction to the theory of games and further notational conventions.)

Definition 4.1 (Game Form)**.** *A* game form *S is a tuple $\langle N, (A_i) \rangle$, where N is a finite set of individual agents, for each agent i in N it holds that A_i is a non-empty and finite set of actions available to agent i. The set of action profiles A is given by $\times_{i \in N} A_i$.*

To study various types of we-intentions, we need to supplement these game forms with intentions. Though philosophers have studied various guises of intentions, I restrict my attention to future-directed intentions as studied in the

[25] When assuming the natural candidate for the group preference in the Hi-Lo game, the theory of team reasoning solves the Hi-Lo game by letting the agents adopt a particular reasoning method, my theory of team-directed intentions, instead, can be interpreted as solving the Hi-Lo game by letting the agents adopt a particular intention.

[26] As noted in Footnote 3, these opposing results derive from the fact that Hakli et al. (2010) assume that I-mode reasoning is connected to equilibrium-based reasoning, whereas I assume that I-mode reasoning is connected to dominance reasoning.

planning theory of intentions advanced by Bratman (1987).[27] There are two different types of future-directed intentions: I can intend to perform a certain action, or I can intend to realize a certain state of affairs. I focus primarily on intentions to realize a certain state of affairs. Because I assume that an action profile fully determines the future state of the world, an intention is then identified with a set of action profiles.[28] Intuitively, an intention $J \subseteq A$ is an intention to realize the aspects that all outcomes of the action profiles in J have in common. For simplicity's sake, I restrict to agents having just a single intention. So the intention of an – individual or collective – agent $\mathcal{H}$ is given by a set of action profiles $Int_{\mathcal{H}} \subseteq A$. This induces the reading that an agent $\mathcal{H}$ intends to φ if and only if her intention $Int_{\mathcal{H}}$ is represented by φ. To model that rational intentions are feasible, I require that $Int_{\mathcal{H}} \neq \emptyset$.

Definition 4.2 (Game with Intentions)**.** *A* game with intentions *S is a tuple $\langle N, (A_i), (Int_{\mathcal{H}})\rangle$, involving a game form $\langle N, (A_i)\rangle$ and intentions $Int_{\mathcal{H}} \in A$ ($Int_{\mathcal{H}} \neq \emptyset$), one for each group of agents $\mathcal{H}$. To picture these intentions, I often use utilities $(u_{\mathcal{H}})$ where*

$$u_{\mathcal{H}}(a) = \begin{cases} 1, \textit{ if } a \in Int_{\mathcal{H}} \\ 0, \textit{ otherwise.}^{29} \end{cases}$$

A game model with intentions *is a game with intentions supplemented with a valuation $V : \mathcal{P} \to 2^A$ (see also § 1.3).*

Intentions provide a "*filter of admissibility* for options" (Bratman, 1987, p. 33). Although Bratman does not use this term in any decision-theoretic sense, I accept

[27] These future-directed intentions (such as my intention to submit this paper by the end of the month) have been distinguished from intentions in action (such as my typing with the intention to finish this introduction) and intentional acts (such as my typing these words intentionally) (see Anscombe (1963)).

[28] As discussed in § 1.3, game forms correspond to *deterministic* STIT models. In this regard, the framework is very similar to that of Van Hees and Roy (2008). Alternatively, if one wants to retain indeterminism, an agent called 'nature' can be added to model the indeterminacy. That is, once every agent has made her choice, the exact outcome is determined by nature's move.

[29] Although my conceptual analysis involves intentions, a decision theorist can view my formal analysis as being restricted to only binary utility functions.

the intuition that an agent intending to φ is required to avoid inadmissible actions, i.e. avoid dominated actions. (See § 1.1 for a more elaborate discussion of this dominance principle.) Admissibility captures the idea that an agent takes all actions of the other agents into consideration; none is entirely ruled out. Avoiding inadmissible actions is more restrictive than avoiding *strictly* dominated actions. I thus take 'providing a filter of admissibility' to mean that an agent's intention requires her to choose an admissible action, that is, one that is not dominated.[30]

In traditional rational choice theory it is common to derive this dominance ordering from exogenously given utilities (see § 1.1). However, I am interested in the dominance orderings resulting from the endogenously adopted intentions. I therefore submit that a dominance ordering is relative to a certain intention. A group action $a_{\mathcal{G}}$ is admissible with respect to an intention to J if and only if no other group action promotes the realization of J, regardless of what the group's non-members do, better than $a_{\mathcal{G}}$ does. So I require that an agent intending to φ should also promote φ, that is, should perform an action that is admissible with respect to φ.

Since I only consider intentions to realize a state of affairs, the adopted dominance ordering is relative to the realization of some state of affairs, instead of maximizing payoffs, as is usually the case in traditional rational choice theory.[31] Moreover, note that altering the state of affairs impacts the resulting dominance ordering. In decision-theoretic terms, an agent may want to optimize her own happiness, but she could instead aim at optimizing their collective, perhaps aver-

[30] Note that I do not employ iterated admissibility, i.e. iterated deletion of dominated actions (see the discussion in Kohlberg and Mertens (1986, Section 2.7) and the discussion of epistemic characterizations of iterated admissibility in Brandenburger et al. (2008, Section 2.6)). One of the main reasons for refraining from doing so is that iterated admissibility is subject to some paradoxes: for instance, it is well known that the order in which dominated strategies are eliminated can affect the outcome of the process. Furthermore, I believe the core results (Results 1 through 4) of this chapter are sustained if iterated admissibility is taken to entail that inadmissible actions are avoided.

[31] However, note that, despite the absence of beliefs and degrees of beliefs, these are intimately related as I interchangeably model the intentions by using a set of possible worlds and a utility function (see Definition 4.2).

age or minimum, happiness; pursuing these different states of affairs may result in different dominance orderings. Still, the principle by which the dominance ordering results from a certain state of affairs is uniform.

The principle guiding the dominance orderings combines the sure-thing principle and reasoning by cases.[32] More specifically, a group action $a_{\mathcal{G}}$ weakly dominates $a'_{\mathcal{G}}$ with respect to an intention if and only if $a_{\mathcal{G}}$ promotes realizing that intention at least as well as $a'_{\mathcal{G}}$, regardless of what the group's non-members do (see § 1.1 and Definitions 1.5 and 2.4).

Definition 4.3 (Dominance)**.** *Let $S = \langle N, (A_i)\rangle$ be a game form, let $\mathcal{G} \subseteq N$ be a group of agents, and let $J \subseteq A$ represent the group's collective intention. Let $a_{\mathcal{G}}$ and $a'_{\mathcal{G}}$ be group actions available to $\mathcal{G}$. Then $a_{\mathcal{G}}$*weakly dominates *$a'_{\mathcal{G}}$* with respect to *J, notation $a_{\mathcal{G}} \succeq_J a'_{\mathcal{G}}$, is defined by:*

$$a_{\mathcal{G}} \succeq_S a'_{\mathcal{G}} \quad \textit{iff} \quad \textit{for all } a''_{-\mathcal{G}} \in A_{-\mathcal{G}} \textit{ it holds that } (a'_{\mathcal{G}}, a''_{-\mathcal{G}}) \in J \textit{ implies } (a_{\mathcal{G}}, a''_{-\mathcal{G}}) \in J.$$

Strong dominance is defined in terms of weak dominance: $a_{\mathcal{G}} \succ_J a'_{\mathcal{G}}$ if and only if $a_{\mathcal{G}} \succeq_J a'_{\mathcal{G}}$ and $a'_{\mathcal{G}} \not\succeq_S a_{\mathcal{G}}$.

The set of *admissible group actions with respect to J* that are available to a group $\mathcal{G}$ in a game form S are defined in terms of the dominance ordering on $A_{\mathcal{G}}$. A group action $a_{\mathcal{G}}$ in $A_{\mathcal{G}}$ is admissible if and only if it is not strongly dominated by any group action in $A_{\mathcal{G}}$:

Definition 4.4 (Admissible Actions)**.** *Let $S = \langle N, (A_i)\rangle$ be a game form, let $\mathcal{G} \subseteq N$ be a group of agents, and let $J \subseteq A$ represent the group's collective intention. Then the set of $\mathcal{H}$'s* admissible actions in *S* with respect to *J, denoted by $Admissible_S(\mathcal{G}, J)$, is given by*

[32] Savage (1954, p. 21) writes: "I know of no other extralogical principle governing decisions that finds such ready acceptance." My personal inspiration is from Horty (1996, 2001), who provided a similar analysis in deontic logic, which is the formal study of obligations and permissions, by introducing "an ordering on actions available to the agent through a state-by-state comparison of their results", where "we will identify the states confronting the agent at any given moment with the possible patterns of actions that might be performed at that moment by all other agents" (Horty, 2001, p. 67 and p. 66).

$$Admissible_S(\mathcal{G}, J) \quad = \quad \{a_{\mathcal{G}} \in A_{\mathcal{G}} \mid \text{there is no } a'_{\mathcal{G}} \in A_{\mathcal{G}} \text{ such that } a'_{\mathcal{G}} \succ_J a_{\mathcal{G}}\}.$$

When we represent the content of an intention by a binary utility function rather than a set of possible worlds, this definition translates to: $a_{\mathcal{G}} \succeq_J a'_{\mathcal{G}}$ if and only if for all $a''_{-\mathcal{G}} \in A_{-\mathcal{G}}$ we have $u_J(a'_{\mathcal{G}}, a''_{-\mathcal{G}}) \geq u_J(a_{\mathcal{G}}, a''_{-\mathcal{G}})$ (where $u_J(a) = 1$ if $a \in J$ and 0 otherwise). This highlights a straightforward connection to the standard weak dominance ordering, which is studied in traditional rational choice theory.

Note that this definition implies that for every group of agents and any collective intention there is at least one admissible group action with respect to that collective intention (because A is assumed to be finite).

To illustrate the dominance ordering and admissibility, consider game S_2 in Figure 4.3. First, observe that there is no connection between personal and group intentions in this particular example. Second, regarding the dominance ordering, since agent i's intention is not realized at (a'_i, a'_j), we can see that for agent i only actions a_i and a''_i are admissible with respect to her intention. For agent j, action a''_j is dominated by a_j with respect to her intention, while a_j and a'_j are incomparable and, hence, both admissible. That is, $a_j \succ_{Int_j} a''_j$, $a_j \not\succeq_{Int_j} a'_j$, and $a'_j \not\succeq_{Int_j} a_j$. For the group $\mathcal{G}$, consisting of agents i and j, the admissible group actions with respect to its collective intention $Int_{\mathcal{G}}$ are (a_i, a_j), (a'_i, a_j), and (a_i, a''_j). Since we accept the intuition that an agent intending to φ should avoid inadmissible actions, this means that agent i can choose a_i and agent j can choose a'_j, with regard to their respective individual intention. It follows that the individual intentions allow (a_i, a'_j), which is inadmissible with respect to the collective intention $Int_{\mathcal{G}}$. The individual intentions hence do not guarantee that a best *group action* is performed, which is not a surprise in this example because of our first observation.

4.3.2 Modal Logic of Agency and Intentionality

In this section I introduce a modal-logical language, which supplements the traditional 'seeing to it that' language, abbreviated to: STIT, with two operators. I

		Agent j		
		a_j	a'_j	a''_j
Agent i	a_i	1,1,1	1,0,0	0,1,1
	a'_i	1,1,1	0,0,0	0,0,0
	a''_i	1,0,0	1,1,0	0,0,0

Figure 4.3: *Game model S_2, where the intentions are represented by a triple of utilities representing i's, j's, and {i, j}'s intention.*

briefly give intuitive readings of these operators before considering the respective formal semantics and providing a more detailed conceptual discussion. The language includes the central operators of the STIT theory of agency, viz. $[\mathcal{H}\ \mathsf{stit}]\varphi$ and $\Box\varphi$, and adds two operators: $[\mathcal{H}\ \mathsf{int}]\varphi$ expresses that 'group $\mathcal{H}$ intends to φ' and $[\mathcal{H}\ \mathsf{prom}]\varphi$ expresses that 'group $\mathcal{H}$ promotes φ regardless of what the others do'. (See § 1.2 for a more detailed introduction to the STIT theory of agency and further notational conventions.)

Definition 4.5 (Syntax)**.** *The formal language $\mathfrak{L}_{ISTIT}$ is as follows:*

$$\varphi ::= p \mid \varphi \wedge \varphi \mid \neg\varphi \mid \Box\varphi \mid [\mathcal{H}\ \mathsf{stit}]\varphi \mid [\mathcal{H}\ \mathsf{int}]\varphi \mid [\mathcal{H}\ \mathsf{prom}]\varphi,$$

where p ranges over a given countable set of propositions $\mathcal{P}$ and $\mathcal{H}$ ranges over subsets of a given finite set of agents N.

The game models with intentions and the dominance ordering, presented in the previous section, are used to provide formal semantics for this logical language:

Definition 4.6 (Semantics)**.** *Let $S = \langle N, (A_i), (Int_{\mathcal{H}}), \pi\rangle$ be a game model with intentions and let $\varphi \in \mathfrak{L}_{ISTIT}$. Then the truth of φ at a profile a in S, notation: $S, a \vDash \varphi$, is given by the following (suppressing the standard propositional clauses):*

$S, a \vDash \Box\varphi$	*iff*	*every* $b \in A$ *satisfies* $S, b \vDash \varphi$;
$S, a \vDash [\mathcal{H}\ \mathsf{stit}]\varphi$	*iff*	*every* $b \in A$ *with* $b_{\mathcal{H}} = a_{\mathcal{H}}$ *satisfies* $S, b \vDash \varphi$;
$S, a \vDash [\mathcal{H}\ \mathsf{prom}]\varphi$	*iff*	$a_{\mathcal{H}}$ *is admissible with respect to* $\{b \mid S, b \vDash \varphi\}$;[33]
$S, a \vDash [\mathcal{H}\ \mathsf{int}]\varphi$	*iff*	*we have* $Int_{\mathcal{H}} = \{b \mid S, b \vDash \varphi\}$.[34]

These semantics implement the idea from STIT theories that 'the agent sees to it that φ' means that the truth of φ is guaranteed by an action or choice of the agent. When Ann empties her glass of milk, the nature of her action on this view is to constrain the possible worlds to those where the glass of milk is emptied. Or, equivalently, the nature of her action is to exclude the possible worlds in which the glass is not emptied. Hence, an action a_i is identified with a subset of the possible worlds, namely those action profiles b satisfying $b_i = a_i$. This induces the reading that an agent sees to it that φ only if she performs an action a_i, thereby constraining the possible worlds to only φ-worlds.[35]

It may be useful to add that the STIT modality $[\mathcal{H}\ \mathsf{stit}]\varphi$ can be interpreted, relative to a profile a, as 'group $\mathcal{H}$ guarantees that φ holds regardless of what the others do'. Indeed, the truth condition of $S, a \vDash [\mathcal{H}\ \mathsf{stit}]\varphi$ is equivalent to requiring that for every $b_{-\mathcal{H}} \in A_{-\mathcal{H}}$ we have $S, (a_{\mathcal{H}}, b_{-\mathcal{H}}) \vDash \varphi$. In other words, in playing $a_{\mathcal{H}}$ group $\mathcal{H}$ is α-effective for φ. This reveals that the group has enough control to ensure φ.

The dual STIT modality $\langle\mathcal{H}\ \mathsf{stit}\rangle\varphi$ expresses that 'group $\mathcal{H}$ allows for φ', that is, group $\mathcal{H}$'s action does not rule out φ. At a profile $a \in A$, this means that there is a combination of actions of the other agents $b_{-\mathcal{H}} \in A_{-\mathcal{H}}$ such that $S, (a_{\mathcal{H}}, b_{-\mathcal{H}}) \vDash \varphi$. Or, equivalently, group action $a_{\mathcal{H}}$ is compatible with φ.

[33] A technical remark: the semantics for the $[\mathcal{H}\ \mathsf{prom}]$ operator could also have been given by neighbourhood semantics, where the neighbourhoods for a group $\mathcal{H}$ of a profile a would be given by the collection $\{M \subseteq A \mid a_{\mathcal{H}}$ is admissible with respect to $M\}$.

[34] Note the equality sign, which means that we employ neighbourhood semantics. If we had employed the standard possible-worlds semantics, it would have read '$Int_{\mathcal{H}}$ consists only of φ-worlds', meaning $Int_{\mathcal{H}} \subseteq \{a \in A \mid a$ satisfies $\varphi\}$. Konolige and Pollack (1993) were the first to model intentions using neighbourhood semantics instead of the standard possible-worlds semantics.

[35] Recall that, in line with the correspondence between STIT models and games (Theorem 1.1), the action profiles in S correspond to possible dynamic worlds in the associated STIT model.

The modality $[\mathcal{H}\ \mathsf{prom}]\varphi$ expresses that group $\mathcal{H}$ performs a best group action with respect to realizing φ, where 'best' is equated with 'admissible'. The underlying intuition is that there may be multiple best actions, each of which is admissible. So, this operator expresses that group $\mathcal{H}$ avoids actions that are inadmissible with respect to φ. This operator hence refers to admissibility, as introduced in Definition 4.4.[36] The intuition that an agent intending to φ is required to avoid actions that are inadmissible with respect to φ thus means that an agent intending to φ is required to promote φ.

One may ask whether the $[\mathcal{H}\ \mathsf{prom}]$-operator can be reduced to an $\mathfrak{L}_{STIT}$-formula.[37] In general, this is impossible (Lemma D.1). This means that extending the standard STIT logic with the $[\mathcal{H}\ \mathsf{prom}]$-operator increases the expressivity of the logic. However, for a given game model S, if there is a group action $a_{\mathcal{H}} \in A_{\mathcal{H}}$ that is $\preceq_\varphi$-maximal in $A_{\mathcal{H}}$, then the $[\mathcal{H}\ \mathsf{prom}]$-operator can be reduced to an $\mathfrak{L}_{STIT}$-formula, viz. $[\mathcal{H}\ \mathsf{prom}]\varphi$ is equivalent to $[\mathcal{H}\ \mathsf{stit}](\langle -\mathcal{H}\ \mathsf{stit}\rangle\varphi \to \varphi)$ in such models (Lemma D.2). The latter formula expresses that group $\mathcal{H}$ guarantees that if φ does not hold, then the other agents guarantee that φ does not hold regardless of what group $\mathcal{H}$ does. Since these details need not distract us at the moment, I refer the interested reader to Appendix D.

The semantics of the intention operator emphasize that we only consider intentions to realize a certain state of affairs, which are represented by a set of worlds. The truth condition for $[\mathcal{H}\ \mathsf{int}]\varphi$ employs neighbourhood semantics and thus induces the reading that the group's intention is represented by φ. These neighbourhood semantics are usually employed to avoid problems concerning logical omniscience, that is, to avoid intentions being closed under logical implications. More specifically, these semantics are typically used to avoid the

[36] A similar operator has been used by Broersen (2011b) to model attempts. Whereas he uses maximizing expected utility, I adopted admissibility as the underlying decision principle. Although there might be some connections to attempts, here I do not want to argue that the $[\mathcal{H}\ \mathsf{prom}]$-operator adequately models attempts.

[37] Recall that $\mathfrak{L}_{STIT}$ only contains the operators $\Box\varphi$ and $[\mathcal{H}\ \mathsf{stit}]\varphi$ (Definition 1.11).

side-effect problem.[38] My motivation is different: it is generally impossible for an agent to perform an action that is both admissible with regard to her intention and admissible with regard to all its logical consequences.[39]

Since my aim is to contribute to a conceptual analysis of collective agency and team reasoning, a complete logical investigation is well beyond my current ambition.[40] Still, it is useful for my current purposes to examine some logical properties of $\mathfrak{L}_{ISTIT}$:

Proposition 4.1. *Let φ be an $\mathfrak{L}_{ISTIT}$-formula and let $\mathcal{H}$ be a group of agents, possibly a singleton. Then*

1. $\vDash \Diamond[\mathcal{H}\ \mathsf{prom}]\varphi$, *one is always able to promote φ,*
2. $\nvDash [\mathcal{H}\ \mathsf{prom}]\varphi \rightarrow [\mathcal{H}\ \mathsf{stit}]\varphi$, *promoting φ does not entail ensuring φ,*
3. $\vDash [\mathcal{H}\ \mathsf{stit}]\varphi \rightarrow [\mathcal{H}\ \mathsf{prom}]\varphi$, *guaranteeing φ entails promoting φ,*
4. $\vDash [\mathcal{H}\ \mathsf{prom}]\varphi \wedge \Diamond\varphi \rightarrow \langle\mathcal{H}\ \mathsf{stit}\rangle\varphi$, *promoting φ while φ is possible entails allowing φ,*
5. $\vDash [\mathcal{H}\ \mathsf{prom}]\varphi \wedge \Diamond[\mathcal{H}\ \mathsf{stit}]\varphi \rightarrow [\mathcal{H}\ \mathsf{stit}]\varphi$, *promoting φ while being able to ensure φ entails guaranteeing φ,*
6. $\vDash \Diamond[\mathcal{H}\ \mathsf{stit}]\varphi \rightarrow ([\mathcal{H}\ \mathsf{prom}]\varphi \leftrightarrow [\mathcal{H}\ \mathsf{stit}]\varphi)$, *if one is able to guarantee φ, then promoting φ is equivalent to ensuring φ.*

Item **1** establishes that one is always able to promote φ, irrespective of its logical form. The fact that this includes infeasible properties, in particular logical inconsistencies, may seem unsatisfactory; however, for infeasible φ it does not

[38] This is why Konolige and Pollack (1993, p. 178) decide to model intentions using neighbourhood semantics1.

[39] To see this, reconsider the game model S_2 in Figure 4.3. We have already noticed that only a_i and a_i'' are admissible with respect to Int_i. However, both fail to be admissible with respect to $Int_i \cup \{(a_i', a_j'), (a_i', a_j'')\}$.

[40] However, recall the correspondence results (§ 1.3), and the completeness result for epistemic STIT theory (§ 3.2).

matter what one does, because one's choice of action does not change the fact that φ will not be realized. Items **2** and **3** show that guaranteeing φ is logically stronger than promoting φ. Item **4** expresses that if φ is feasible, promoting φ entails that one performs an action that is compatible with φ. Or, equivalently, if one performs an action that is incompatible with φ, then one is surely not promoting φ. Items **5** and **6** show that, although promoting φ is logically weaker than guaranteeing φ, if one is able to ensure φ, then promoting φ is equivalent to guaranteeing φ. This shows that one is definitely devoted to realizing φ if one promotes φ.

Because my primary interest is in how intentions constrain the choice of strategy of the respective agents, I refrain from a logical analysis of the intention operator. Despite this void, I will investigate the outcomes that certain individual intentions guarantee. To do so, I rely on the intuition that an agent intending to φ is required to avoid actions that are inadmissible with respect to φ. That is, an agent intending to φ is required to promote φ. Individual intentions therefore guarantee that individual agents choose actions that are admissible with regard to their respective intentions. To generalize, a set of individual intentions guarantee a certain property ψ only if it is the case that whenever the individual agents choose actions that are admissible with respect to their intentions, then ψ holds. It is natural to interpret this as a conditional: if agent i intends to φ, and therefore performs an individual action that is admissible with respect to φ, then ψ will hold. That is, whenever the agent promotes φ then ψ will hold.

Definition 4.7. *Let S be a game model with intentions, let $a \in A$ be a profile, let φ and ψ be formulas in $\mathfrak{L}_{ISTIT}$, and let i be an individual agent. Then we say that, in S, $[i\ \mathsf{int}]\varphi$ guarantees ψ if and only if $S, a \vDash \Box([i\ \mathsf{prom}]\varphi \rightarrow \psi)$. (Note that we do not assume $S, a \vDash [i\ \mathsf{int}]\varphi$.)*

More generally, let $\mathcal{G}$ be a group of agents and let ψ and φ_i be formulas in $\mathfrak{L}_{ISTIT}$, one for each $i \in \mathcal{G}$. Then we say that, in S, $([i\ \mathsf{int}]\varphi_i)_{i\in\mathcal{G}}$ guarantee ψ if and only if $S, a \vDash \Box(\bigwedge_{i\in\mathcal{G}}[i\ \mathsf{prom}]\varphi_i \rightarrow \psi)$.

The latter can be semantically interpreted as follows: let $P \subseteq A$ and J_i represent a state of affairs, one for each $i \in \mathcal{G}$. Then we say that, in S, the individual intentions $\{J_i\}_{i\in\mathcal{G}}$ guarantee that P holds if and only if for every $b \in A$, if for every $i \in \mathcal{G}$ it is the case that b_i is admissible with respect to J_i, then property P holds at b, that is, $b \in P$.

The focus, later in the chapter, is going to be on whether certain we-intentions guarantee that a best group action is performed.

4.4 Three Types of We-intentions

Which individual attitudes are warranted in the context of a collective intention? This section serves four purposes. First, in the following, the logical language is used to formalize and study so-called pro-group intentions, team-directed intentions, and participatory intentions. The running example to illustrate these types of we-intentions is displayed in Figure 4.4. Second, in § 4.4.2 I show that team reasoning and team-directed intentions yield the same action recommendations (Result 1). Third, at the end of § 4.4.3 I show, using the running example, that these three types of we-intentions yield different action recommendations. Finally, in § 4.4.4 I briefly elaborate on the implications for standard game-theoretical studies of cooperation and fairness.

I henceforth presuppose a collective intention to φ and investigate what team reasoning and the we-intention types amount to. Just as an individual agent's intention requires her to perform an individual action that is admissible with respect to that individual intention, a collective intention requires the group to perform a group action that is admissible with respect to its collective intention. Or, equivalently, the collective intention provides a filter of admissibility for the available group actions in which the admissible group actions are best. The group should therefore perform a group action that promotes the realization of what is collectively intended.

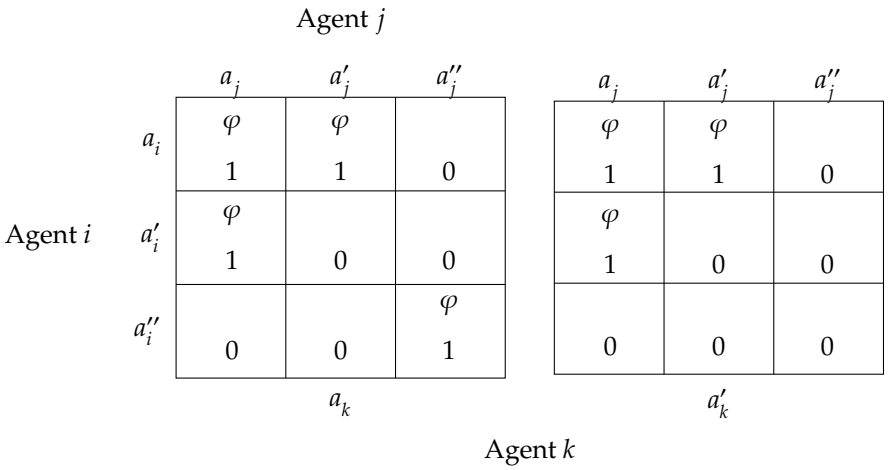


Figure 4.4: *Game model S_4: a three-player game, where $\{i, j\} = \mathcal{G}$ collectively intends to φ, showing only the collective intention of $\mathcal{G}$.*

4.4.1 Pro-group Intentions

A member i of $\mathcal{G}$ could adopt the collective goal as her own, instead of furthering her personal goals, and pursue it to the best of her abilities, expressed by $[i\ \mathsf{int}]\varphi$ and coined a *pro-group intention*. This way, she ignores the contributions others can make and does her best to realize the collective intention regardless of what others do. Various game-theoretic enterprises try to explain cooperative behaviour by transforming the preferences of the group members (see § 4.4.4); the same intuition underlies this first we-intention type:

Definition 4.8 (Pro-group Intentions)**.** *Suppose group $\mathcal{G}$ collectively intends to φ. Let $i \in \mathcal{G}$ be a member of the group $\mathcal{G}$. Agent i's* pro-group intention *is an intention to promote the group's objective, which is expressed by* $[i\ \mathsf{int}]\varphi$.

In game model S_4 of Figure 4.4, we observe that agents i and j would have the pro-group intention if their intentions were represented by φ, as depicted in Figure 4.5. Let us consider which individual actions would be admissible, that

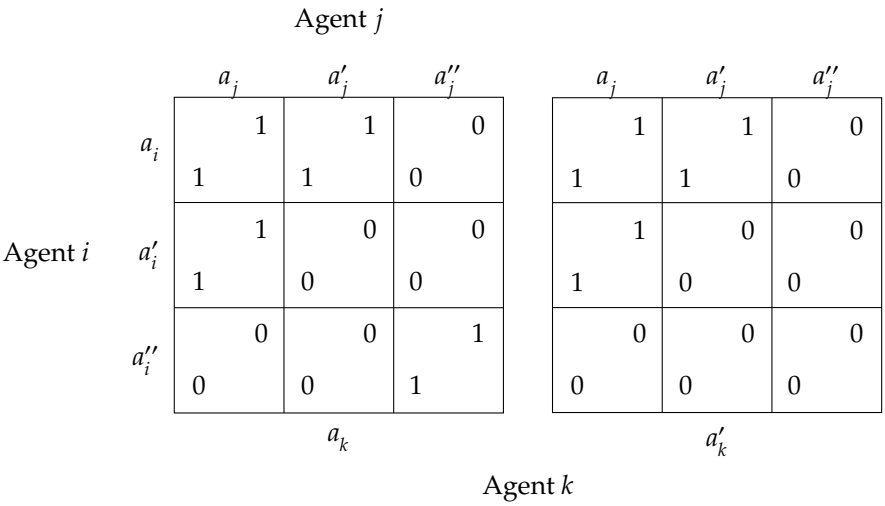


Figure 4.5: *The pro-group intentions of individual agents i and j in game model S_4.*

is, not dominated, if the agents were to adopt the pro-group intention: we have $a_i \succ_\varphi a'_i$, because for instance $S_4, (a'_i, a'_j, a_k) \nvDash \varphi$ and $S_4, (a_i, a'_j, a_k) \vDash \varphi$,. And, because $S_4, (a''_i, a'_j, a_k) \nvDash \varphi$, we also have $a''_i \not\succeq_\varphi a_i$. Likewise, by comparing (a_i, a''_j, a_k) and (a''_i, a''_j, a_k), we derive that $a_i \not\succeq_\varphi a''_i$. Hence, if agent i adopted the pro-group intention, only a_i and a''_i would be admissible with respect to her intention. By symmetry, if agent j adopted the pro-group intention, only a_j and a''_j would be admissible with respect to her intention.

4.4.2 Team-directed Intentions

Discontent with preference transformations, Bacharach, Sugden, and Gold argue, on various occasions, that team reasoning is more appropriate for explaining and predicting cooperative behaviour. Instead, they propose an *agency* transforma-

tion. When engaging in team reasoning, a member of a group first identifies a best combination of individual actions that the group members can perform and then decides to perform the individual action that is her part of that combination.

To illustrate the benefit of team reasoning in games with intentions, consider the *alternative Hi-Lo game* depicted in Figure 4.6. Note that, in the context of group $\mathcal{G}$'s collective intention to φ, the best group action is (*high*, *high*), because it is the only group action that ensures φ. An agent engaged in team reasoning therefore first identifies (*high*, *high*) as the unique best group action and then decides to perform her part in that combination, therefore recommending *high* to both agent i and agent j. So, in this game, team reasoning by individual agents i and j ensures that group $\mathcal{G}$ performs a best group action, that is, a group action that is admissible with respect to its collective intention.

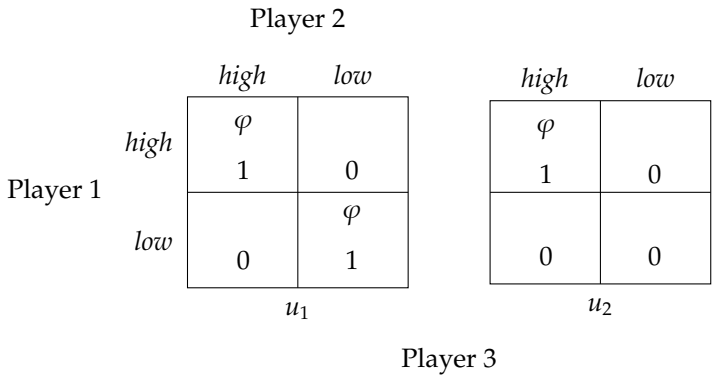


Figure 4.6: *The alternative Hi-Lo game S_3: a three-player game, where {player 1, player 2} = $\mathcal{G}$ collectively intends to φ, showing only the collective intention of $\mathcal{G}$.*

To show that team reasoning yields the same action recommendations as a certain we-intention, I introduce *team-directed intentions*, which are standard individual intentions with a certain type of content. An agent adopting a team-directed intention transforms her preferences and adopts as her personal objective

performing an individual action that is compatible with a best group action. This may require her to put herself in the shoes of the group agent to determine the best group actions. Then she takes as her personal objective performing an individual action that is compatible with a best group action and decides what to do by way of individualistic reasoning. The realization of her objective hence does not depend the actions of other agents: any action profile in which she performs an individual action that is compatible with a best group action realizes the team-directed intention. To find the best group actions, she applies the dominance principle at the level of the group. Then, she decides what to do by determining the individual actions that are admissible with respect to her team-directed intention, which are exactly those individual actions that are compatible with a best group action.

Definition 4.9 (Team-directed Intentions)**.** *Suppose group $\mathcal{G}$ collectively intends to φ. Let $i \in \mathcal{G}$ be a member of the group $\mathcal{G}$. Individual agent i's* team-directed intention *is an intention to act in a way that is compatible with a best group action, which is expressed by* $[i\ \mathsf{int}]\langle i\ \mathsf{stit}\rangle[\mathcal{G}\ \mathsf{prom}]\varphi$.

It may be helpful to note that team-directed intentions differ from pro-group intentions. After all, if the group faces uncertainty then the collective goal could be realized although the individual agent ensures that the group does not perform a best group action. For example, in the alternative Hi-Lo game, the collective goal is realized in (*low*, *low*, u_1) even though player 1 rules out that a best group action, viz. (*high*, *high*), is realized if she chooses *low*. So in (*low*, *low*, u_1) the pro-group intention is realized while the team-directed intention is not.

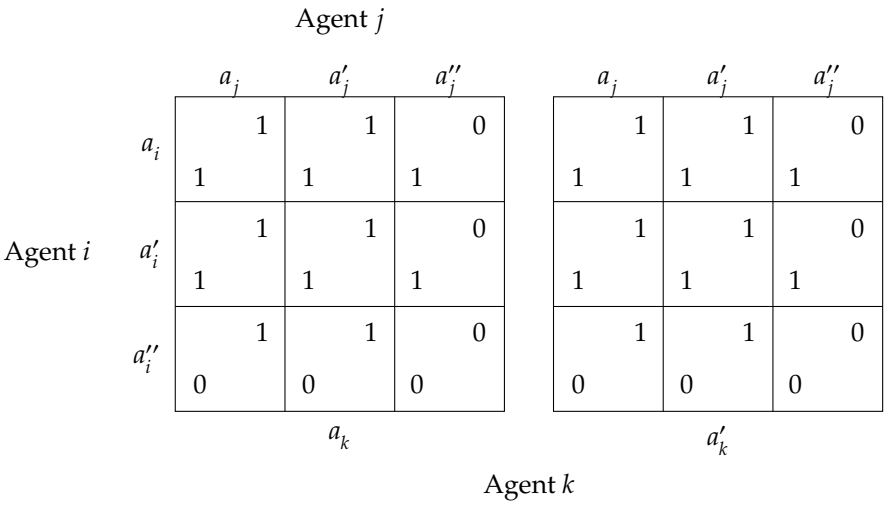
To get some feeling for the recommendations yielded by team-directed intentions, reconsider game model S_4 of Figure 4.4. When the members would adopt the team-directed intentions in S_4 then the scenario would be depicted by Figure 4.7a. Let me explain why. First, note that, for the group $\mathcal{G}$ consisting of agents i and j, (a_i, a_j), (a_i', a_j), and (a_i, a_j') are the only group actions that are admissible with

respect to φ. Indeed, since these are the only group actions that ensure that φ is realized, these are the best group actions (see, for instance, Proposition 4.1 item **6**). So, for agent i, for example, any profile in which she performs a_i or a'_i will realize her team-directed intention. It then follows that only a_i and a'_i are admissible with respect to agent i's team-directed intention. By symmetry, if agent j adopted the team-directed intention, only a_j and a'_j would be admissible with respect to agent j's team-directed intention.

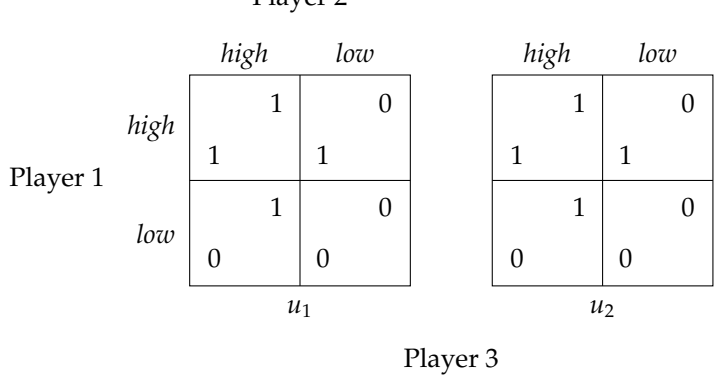
An individual agent performs an action that is admissible with respect to her team-directed intention if and only if she performs an individual action that is her component of a best combination of actions that the members can perform (Observation D.1). Although team-reasoning theorists typically presuppose that there is a *unique* best group action, I follow Bacharach (1999, p. 120 – adapted notation): "I shall say that agent i team reasons (for $\mathcal{G}$) when she first computes a best profile $a^*_{\mathcal{G}}$ (in terms of $u_{\mathcal{G}}$), next computes a^*_i, and lastly decides to do a^*_i because this is the component under her control of a best profile." It is unsurprising that individual actions resulting from team reasoning coincide with individual actions admissible with respect to the team-directed intention. (All claims are proved in Appendix D.)

Result 1. *Let $S = \langle N, (A_i), (Int_{\mathcal{H}}), \pi\rangle$ be a game model. Suppose group $\mathcal{G}$ collectively intends to φ. Let $a \in A$. Then team reasoning admits the individual action a_i if and only if $S, a \vDash \langle i \ \mathsf{stit}\rangle[\mathcal{G} \ \mathsf{prom}]\varphi$, which is, in turn, equivalent to saying that the individual action a_i is admitted by the team-directed intention.*

This result shows that the results of team reasoning can be equally well explained by team-directed intentions. Or, equivalently, it shows that an important part of we-mode reasoning can be reduced to I-mode reasoning with a team-directed intention. So, the effects of the agency transformation in team reasoning can be mirrored by the preference transformation in team-directed intentions.



(a) *The team-directed intentions of individual agents i and j in the game model S_4.*



(b) *The team-directed intentions of players 1 and 2 in the alternative Hi-Lo game S_3.*

Figure 4.7

In particular, this means that team-directed intentions provide the same action recommendations as team reasoning in the alternative Hi-Lo game S_3 of Figure 4.6. Note that (*high*, *high*) is the only group action of $\mathcal{G}$ that ensures φ regardless of what player 3 does. Therefore, for example, player 1's team-directed intention is only realized in profiles where her component is *high*. Hence, Figure 4.7b depicts S_3 when $\mathcal{G}$'s members would adopt the team-directed intention.

This connection shows that we need not focus on the *mental processes* by which collective intentions are formed, because it suffices to study the *mental states* of the members.[41] The result therefore complements the analysis by Gold and Sugden (2007), who argue that team reasoning leads to collective intentions:

> Team reasoning results in the formation of intentions. ... references to the group are noneliminable parts of the reasoning process that led to the formation of the intention. Thus, it is natural to regard the intentions that result from team reasoning as collective intentions. (Gold and Sugden, 2007, p. 126)

If all of this is correct, then the connection between team-directed intentions and team reasoning, established by the result, reveals that it is equally natural to suppose that team-directed *intentions* are essential for collective intentions. In line with the philosophical literature, this purports a relation between personal and collective intentionality.

4.4.3 Participatory Intentions

Now we direct our attention to participatory intentions. Observing that team reasoning corresponds to we-intentions of the form $[i\ \mathsf{int}]\langle i\ \mathsf{stit}\rangle[\mathcal{G}\ \mathsf{prom}]\varphi$ invites

[41] Compare Hakli et al. (2010, p. 299): "We study processes of we-reasoning concentrating on the difference between we-mode reasoning and pro-group I-mode reasoning. The difference is not in the aims of the agents because in both cases the agents aim at the benefit of the group. Rather, the difference is in the reasoning process: It is individualistic in the I-mode case."

a natural objection. This objection originates from the oddity of using three consecutive modalities to express a team-directed intention. Team-directed intentions merely require the members to perform an individual action that is *only compatible* with a best group action. This is a very weak demand. Therefore, we introduce a third we-intention type: *participatory intentions*, which require an individual agent to *promote* the realization of a best group action. That is, she adopts as her objective that a best group action is performed and then decides what to do by way of individualistic reasoning.

Definition 4.10 (Participatory Intentions)**.** *Suppose group $\mathcal{G}$ collectively intends to φ. Let $i \in \mathcal{G}$ be a member of the group $\mathcal{G}$. Agent i's* participatory intention *is an intention that the group promotes the group's objective, which is to realize φ, regardless of what others do, that is,* $[i\ \mathsf{int}][\mathcal{G}\ \mathsf{prom}]\varphi$.

Note the minor, yet crucial, difference: pro-group intentions require the individual agent to adopt the *group's objectives* as her own, whereas a participatory intention requires her to adopt as her personal objective that a *best group action is performed*. This reveals that participatory intentions can be viewed as a preference transformation. This preference transformation is not the result of the aggregation of preferences, but crucially relies on group notions. After all, an agent forming such a participatory intention is required to answer the question 'What should *we* do?' before being able to answer the question 'What should *I* do (with respect to my participatory intention)?'. The appeal to an agency transformation in the team-reasoning literature is thus taken up by participatory intentions. However, in contrast to altering the reasoning process, a participatory intention alters the preferences.

It may be helpful to note that participatory intentions differ from team-directed intentions. After all, an individual agent may perform an individual action that is compatible with a best group action even though no best group action is performed. For example, in the alternative Hi-Lo game, at ($high$, low, u_1) the best

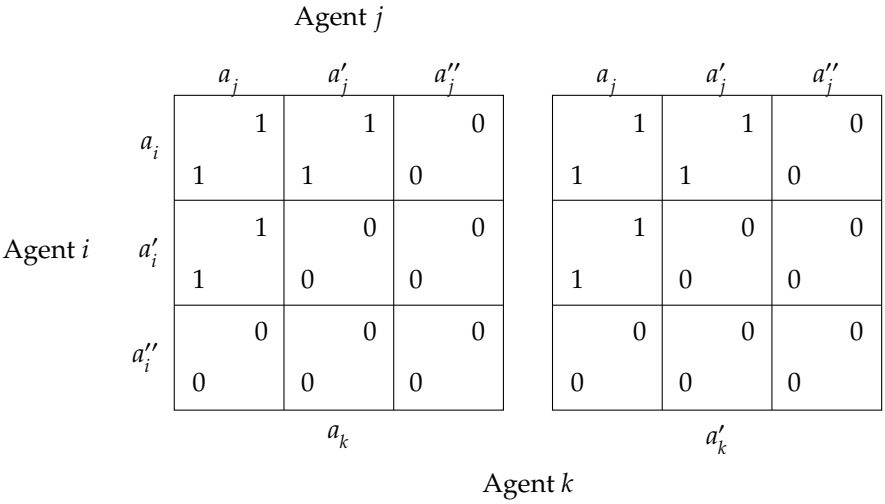
group action, viz. (*high*, *high*), is not performed despite the fact that player 1 performs an individual action that is compatible with the best group action. So in (*high*, *low*, u_1) player 1's team-directed intention is realized while her participatory intention is not.

Reconsider game model S_4 of Figure 4.4. Again, note that (a_i, a_j), (a'_i, a_j), and (a_i, a'_j) are the only group actions that are admissible with respect to φ for the group $\mathcal{G}$ consisting of agents i and j. This means that Figure 4.8a accurately represents the scenario when the members would adopt the participatory intention. It follows that the individual action a''_j is incompatible with a best group action, and that a_j dominates a'_j. Therefore, only a_j is admissible with respect to agent j's participatory intention. By symmetry, if agent i adopted the participatory intention, we derive that only a_i is admissible with respect to agent i's participatory intention. It is easy to see that Figure 4.8b presents the alternative Hi-Lo game if the members would adopt the participatory intention.

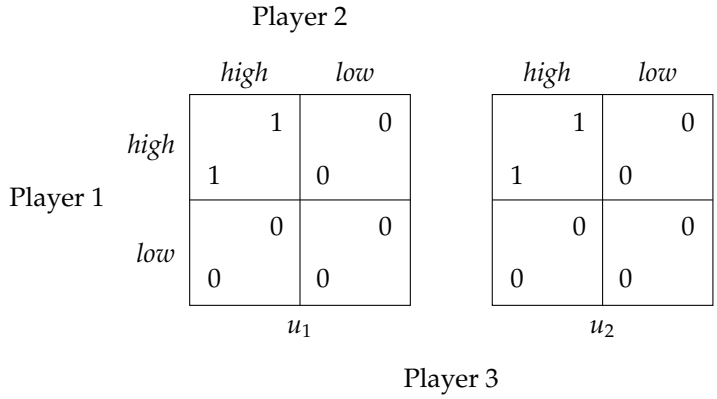
In this section, I have introduced and discussed three we-intention types, but why bother? The discussions of game model S_4 illustrate that these three we-intention types yield different action recommendations. If agent i adopted the pro-group intention, then individual actions a_i and a''_i would be admissible. If instead she adopted the team-directed intention, then individual actions a_i and a'_i would be admissible. Finally, only individual action a_i would be admissible if she adopted the participatory intention. This shows that there is a significant difference between these three we-intention types.

4.4.4 Fairness and Cooperation

At this point it may be useful to explain in some detail how these we-intentions can be contrasted with the existing literature on fairness and cooperation in game theory. More specifically, it might be helpful to explicate why these we-intentions



(a) *The participatory intentions of i and j in the game model* S_4.



(b) *The participatory intentions of players 1 and 2 in the alternative Hi-Lo game* S_3.

Figure 4.8

can be viewed as an *extension* of preference transformation theories. This is important because the team-reasoning literature has made two mutually reinforcing claims regarding preference transformations:

1. It is claimed that team reasoning cannot be captured by a preference transformation. Rather, team reasoning is best viewed as an *agency transformation*.

2. It is claimed that preference transformations cannot explain why anyone should play *high* in the Hi-Lo game, nor can they explain why most people actually choose *high*.

This failure points to a serious restriction of preference transformation theories, one that the team-reasoning account of cooperation aims to overcome. To explain, and eventually *discount*, these two claims, it is helpful to start by discussing some prominent preference transformation theories.

It is well known that standard game theory is unable to explain some trivial examples of human decision-making. For instance, the traditional theory of self-interested rational individuals cannot explain, at least not satisfactorily, why people vote, pay their taxes, or sacrifice their own prospects in favour of those of a peer. To address this lacuna, Matthew Rabin believes we need a model that incorporates three facts:

> (A) People are willing to sacrifice their own material well-being to help those who are being kind.
>
> (B) People are willing to sacrifice their own material well-being to punish those who are being unkind.
>
> (C) Both motivations (A) and (B) have a greater effect on behavior as the material cost of sacrificing becomes smaller. (Rabin, 1993, p. 1282)

Rabin develops a theory that is best interpreted as psychological game theory, which is the field of game theory that aims to provide a theory that explains

human behaviour and experimental findings rather than a theory that provides normative guidance for agents that face interdependent decision problems. Rabin gives a method to transform a so-called material game into a psychological game by amending the utilities. This means that Rabin's theory of fairness is best viewed as a preference transformation theory. The *expected transformed utility* for an individual agent will depend on three factors: "(i) his strategy, (ii) his beliefs about the other player's strategy choice, and (iii) his beliefs about the other player's beliefs about his strategy" (Rabin, 1993, p. 1286). The details need not worry us at present, but it may be helpful to note that his theory relies on so-called kindness functions to model this transition. These and similar models have been used to fruitfully explain various empirical findings (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000).[42]

What is the relation between we-intentions and preference transformation theories? It is important to note that preference transformation theories use rationality principles from orthodox rational choice theory to derive the predicted or observed behaviour in experimental settings. That is, these theories *only* change the utilities; a decision-maker then aims to maximize expected transformed utility. If we assume that these preference transformation theories accurately model the group utility, then a preference transformation theory is best interpreted using pro-group intentions: the members of the group adopt the transformed preference as their own and then reason individualistically.[43] Accordingly, a preference transformation theory may be implemented in my theory of we-intentions to yield predicted or observed behaviour in experimental settings.

[42]To illustrate the popularity of this methodological strategy, note that Bicchieri (2006, p. 3) studies social norms as a type of preference transformation: "social norms, as I shall argue, *transform* mixed-motive games into coordination ones. This transformation, however, hinges on each individual expecting enough other people to follow the norm, too. If this expectation is violated, an individual will revert to playing the original game and to behaving 'selfishly.'"

[43]Strictly speaking, my theory relies on the admissibility principle rather than expected utility theory. Nonetheless, both are principles from orthodox rational choice theory.

What about team-directed intentions? The theory of team reasoning, in contrast to preference transformation theories, revises the rationality principles of traditional rational choice theory by adopting a new reasoning method. Since the action recommendations yielded by team-directed intentions align with those resulting from team reasoning, there is no behavioural difference between my theory of team-directed intentions and the theory of team reasoning. This means that any experimental evidence for team reasoning will also be experimental evidence for my theory of team-directed intentions.[44] To distinguish between these theories we thus need to go beyond the behavioural realm and include mental constructs.

In light of the two claims made by team-reasoning theorists, it is useful to add that team-directed intentions can be viewed as a preference transformation. An agent adopting a team-directed intention transforms her preferences and adopts as her personal objective performing an individual action that is compatible with a best group action. Then she decides what to do by way of individualistic reasoning. This is certainly a very strange preference transformation and one that might not cohere with other mental constructs in an agent's psychological economy. For example, the transformed preference suggests that an individual agent prefers any outcome in which she herself acts compatibly with a joint act that promotes the realization of the group's objective over any outcome in which she acts differently. Such properties are not standardly included in the outcomes of a game. For instance, in his seminal contribution to decision theory, Savage (1954) models actions as functions from states to outcomes. So standard preferences on outcomes cannot encompass this preference transformation. Following Dietrich

[44] Bardsley et al. (2010) and Faillo et al. (2013) discuss experimental evidence for team reasoning.

and List (2016b), we may say that the preference transformation corresponding to team-directed intentions is based on "context-related" properties.[45] We therefore need not revise the rationality principles of traditional rational choice theory.

The team-directed intentions are remarkable in that the corresponding transformed preference only depends on the agent's own act. The team-directed intention is realized if and only if the agent performs an action that is compatible with a best group action. This may not accord with our intuitions: it would imply that both (*high*, *high*) and (*high*, *low*) best satisfy player 1's preferences (see Figure 4.2). Things are different for participatory intentions: their satisfaction not only depends on an agent's own act but also on what others do. In the Hi-Lo game, an agent who adopts the participatory intention can be taken to prefer an outcome in which everybody plays her part in (*high*, *high*) over any outcome in which someone deviates from playing her part. For these reasons, one could argue that participatory intentions better accord with our intuitions, despite the fact that participatory intentions and team-directed intentions yield the same behavioural predictions in the Hi-Lo game.

In this brief exposition I have shown how we-intentions relate to preference transformation theories and team reasoning. Preference transformation theories are best interpreted using pro-group intentions. I have argued that the two claims of team-reasoning theorists cannot be upheld, at least not without turning to a discussion of mental constructs rather than rational play. After all, the preference transformation corresponding to team-directed intentions can fully account for the behaviour predicted and observed by team reasoning. Nonetheless, team-directed intentions may not accord with our intuitions regarding mental con-

[45] Dietrich and List (2016b, p. 177) write that "the motivationally salient properties may go beyond 'intrinsic' properties of the options and include 'context-related' properties. Examples are whether an option conforms to a context-specific social norm (e.g. is it polite?), whether it is above average quality among the available options, or whether the choice menu offers luxury options".

structs. I pointed out that participatory intentions correct some of the conceptual flaws in team-directed intentions and therefore seem to be the better candidate for future research.

4.5 Participatory Intentions Prevail

The three we-intention types – pro-group, team-directed, and participatory intentions (§ 4.4) – have been introduced to predict and explain cooperative behaviour and incentives. Here I focus on their *effectiveness* with regard to guaranteeing that a best group action is performed. If these intention types have anything to say about cooperation they should certainly advance best group actions across a wide range of games.[46] I thus investigate the classes of games for which these we-intention types guarantee that a best group action is performed. The results of this section are collected in Figure 4.1 on page 150. As the title of this section already reveals, I prove that the theory of participatory intentions, in contrast with team-directed intentions or pro-group intentions, is the *prevalent* account of cooperation.

4.5.1 A Stand-off

It may be thought that even though team-directed intentions and pro-group intentions yield different action recommendations, they nonetheless yield the same cooperative outcomes. In this section, I show that the alternative Hi-Lo game illustrates that this is false. Since team reasoning succeeds in selecting a best group action in this particular game whereas pro-group intentions fail to do so, it might then be thought that team-directed intentions surpass pro-group intentions

[46] Bacharach (2006, p. 58 – adapted notation) writes: "There are three requirements for a good theory of why people play *high* in Hi-Lo: [. . .] (iii) that it be part of a unified theory of a wide range of problems, not just Hi-Lo–for example, all problems of cooperation."

		Agent j: L	Agent j: R
Agent i: U		φ 1	φ 1
Agent i: D		φ 1	 0

Figure 4.9: *Game model S_5, where the group $\{i, j\} = \mathcal{G}$ collectively intends to φ, showing only the collective intention of $\mathcal{G}$.*

with regard to guaranteeing successful cooperation in *all* scenarios. I show that this is also false. This means that team reasoning is on an unsatisfactory par with pro-group intentions.

The alternative Hi-Lo game S_3 of Figure 4.6 on page 175 presents a scenario in which team-directed intentions ensure successful cooperation. Pro-group intentions do not fare well in that example: since neither individual action dominates the other with respect to the group's objective, both are admissible. Pro-group intentions hence fail to dismiss all inferior group actions, for instance (*high*, *low*). I take this to reveal that there are games in which pro-group intentions fail to explain the obvious collective incentives, whereas team reasoning succeeds in doing so.

There are, however, games revealing the opposite. In game model S_5 of Figure 4.9, team-directed intentions do not have much to offer: since group $\mathcal{G}$'s actions (U, L), (U, R), and (D, L) ensure φ, they are the best group actions in the context of the collective intention to φ.[47] So any individual action is compatible with a best group action. In particular, team-directed intentions admit D and R for agents i and j respectively, and hence do not dismiss (D, R). Pro-group intentions do not suffer from this flaw: they recommend U over D and L over R for agents

[47] Note that this game resembles the many-hands problem discussed in § 1.1, Figure 1.2.

i and j, respectively. So, whereas team-directed intentions risk performing an inferior group action, namely (D, R), pro-group intentions guarantee that a best group action is performed, namely (U, L). This illustrates that there is a game in which team reasoning fails to select a best group action whereas pro-group intentions succeed in doing so.

I agree with Bacharach (2006, p. 60, § 8.3) that a theory should only be determinate when our intuitions are. We should hence not presuppose the determinacy of reason.[48] The team reasoning literature has, however, predominantly focused on cases where there is a *unique* best group action.[49] I wish to avoid this restriction and introduce a more general theory of cooperation.[50] When dropping the uniqueness assumption, the above discussion reveals that team reasoning does not surpass pro-group intentions.

My discussion seems to contradict the result of Bacharach (1999, Theorem 2), which is interpreted as showing that "team reasoning differs from, and is more powerful than, adopting the group's objective and then reasoning in the standard individualistic way" (p. 144).[51] Our results are, however, mutually consistent. The difference between my approach and Bacharach's is the employed individualistic reasoning method: I adopt the admissibility requirement, which states that

[48] In contrast, Harsanyi and Selten (1988, p. 13) write: "Clearly a theory telling us no more than that the outcome can be any one of these equilibrium points will not give us much useful information. We need a theory selecting one equilibrium point as the solution of the game."

[49] As mentioned before, Bacharach (1999, p. 120 – adapted notation) is a notable exception: "I shall say that agent i team reasons (for $\mathcal{G}$) when she first computes a best profile a^* (in terms of $u_{\mathcal{G}}$), next computes a_i^*, and lastly decides to do a_i^* because this is the component under her control of a best profile." Also see the helpful discussion on cases where there are multiple best group actions by Karpus and Radzvilas (2018, § 4.7).

[50] Our spirit in Tamminga and Duijf (2017, pp. 187 and 207) is similar, where we "study a strong sense of joint action in which members of a group, using team reasoning, design and then publicly adopt a group plan". Since such a group plan can be indeterminate, our focus is on "structural conditions that a group plan must meet in order to successfully coordinate the individual actions of the group members". See also Chapter 2.

[51] Hakli et al. (2010, Thesis (3), p. 306) agree with Bacharach and write: "The we-mode tends to create more collective order than the pro-group I-mode: It can decrease the amount of equilibria but it cannot increase them."

individuals reason in such a way as to avoid dominated actions; Bacharach relies on equilibrium reasoning, which states that individuals determine an "individualistic best reply" (p. 127).

In sum, I argued that the class of games for which team-directed intentions guarantee that a best group action is performed is not a proper superset of the corresponding class for pro-group intentions, and vice versa. So it is unclear which of these theories offers the best account of cooperation. This results in a stalemate with regard to the range of problems for which they guarantee successful cooperation.

Result 2. *Team-directed intentions and pro-group intentions are on a par with regard to guaranteeing successful cooperation.*

4.5.2 Overcoming the Deadlock

Given the stand-off between team-directed and pro-group intentions, it is natural and important to enquire whether this deadlock can be overcome. In this section, I thus investigate two issues: do participatory intentions guarantee that a best group action is performed when team-directed intentions do so? And when pro-group intentions do so? The answer to both questions turns out to be affirmative. This means that the theory of participatory intentions is the prevalent account of cooperation. That is, if a group of individuals comes together to strive for some collective or joint goal, the best they can do is to each adopt the participatory intention (and act accordingly), rather than adopting a team-directed or pro-group intention.

Let us start with participatory intentions and team-directed intentions. It seems uncontroversial to claim that whenever an individual agent decides to perform an individual action that is incompatible with any best group action, then she is certainly not promoting the realization of a best group action. It indeed follows logically that

$$\vDash [i\ \mathsf{prom}][\mathcal{G}\ \mathsf{prom}]\varphi \to [i\ \mathsf{prom}]\langle i\ \mathsf{stit}\rangle[\mathcal{G}\ \mathsf{prom}]\varphi \quad \text{(Observation D.2).}$$

Participatory intentions hence refine team-directed intentions' action recommendations.

Do participatory intentions guarantee that a best group action is performed if team-directed intentions do? Suppose a scenario is given in which the following holds: when the agents adopt the team-directed intention and therefore perform an individual action that is admissible with respect to it, then ψ will hold. Since participatory intentions refine the action recommendations yielded by team-directed intentions, this means that if the agents adopted the participatory intention, then ψ will also hold. This immediately implies that whenever team-directed intentions guarantee that a best group action is performed, then so do participatory intentions. In particular, this shows that participatory intentions guarantee that a best group action is performed in scenarios in which there is a unique best group action, since team-directed intentions are effective in those scenarios.

Let us briefly pause here. One of the main justifications for team reasoning is that it advances cooperative behaviour in Hi-Lo games. Whether the cooperative incentives in the Hi-Lo game actually lead to team reasoning is not at issue. Team reasoning sets out to address what makes (*high*, *high*) the only rational option. As such, the theory of participatory intentions is on an equal footing. After all, the action recommendations resulting from team reasoning coincide with those resulting from participatory intentions in the Hi-Lo game. This justification for the team-reasoning account of cooperation therefore transfers to the theory of participatory intentions.

Result 3 (below) can be viewed as generalizing and strengthening this justification for the theory of participatory intentions. On a positive note, my result establishes that whenever team reasoning is successful in picking out a best group action, then so are participatory intentions. On a negative note, there are scen-

arios in which team reasoning fails in this respect, whereas participatory intentions succeed.[52] Hence, the team-reasoning account of cooperation is surpassed by the theory of participatory intentions with regard to guaranteeing successful cooperation.

Result 3. *Participatory intentions surpass team-directed intentions with regard to guaranteeing successful cooperation.*

Let us now turn to participatory intentions and pro-group intentions. The stand-off between team reasoning and pro-group intentions originates from game S_5 in Figure 4.9. So let us investigate how participatory intentions fare in that game. Recall that, for group $\mathcal{G}$, group actions (U, L), (U, R), and (D, L) are the best group actions in the context of the collective intention to φ. Since these coincide with the φ-worlds, the results of pro-group intentions and participatory intentions coincide in this game. That is, the action recommendations yielded by these we-intention types are identical in this game. In particular, this shows that participatory intentions champion the objection posed to team reasoning.

Can we come up with a different objection against participatory intentions, in favour of pro-group intentions? That is, does a scenario exist in which participatory intentions fail to guarantee that a best group action is performed whereas pro-group intentions succeed in doing so? It is tempting to think that the action recommendations yielded by participatory intentions refine those of pro-group intentions. This is, however, *not* the case.[53] Still, the answer to the questions is negative: the following result shows that participatory intentions also surpass pro-group intentions with regard to guaranteeing successful cooperation.[54]

[52] The discussion below shows that the game in Figure 4.9 is a case in point.

[53] To see this, imagine if we slightly change game S_4 in Figure 4.4: drop φ at (a_i, a'_j, a'_k) and thus remove (a_i, a'_j, a'_k) from $Int_{\mathcal{G}}$. One can check that if agent i adopted the pro-group intention, only a_i and a''_i would be admissible. In contrast, if agent i adopted the participatory intention, only a_i and a'_i would be admissible. This example hence proves the point.

[54] Because the action recommendations of participatory intentions do not refine those of pro-group intentions, the proof that participatory intentions surpass pro-group intentions is more complicated than that of Result 3.

Result 4. *Participatory intentions surpass pro-group intentions with regard to guaranteeing successful cooperation.*

This result emphasizes that the theory of participatory intentions surpasses team reasoning *and* pro-group intentions in selecting cooperatively rational solutions. Indeed, in *any* scenario in which pro-group intentions guarantee successful cooperation, participatory intentions do so too. So, to cooperate successfully, it is generally better if all members take up the participatory intention.

Considered together, the results of this section purport that the theory of participatory intentions, in contrast with team reasoning or pro-group intentions, is the prevalent account of cooperation. Since team reasoning does not surpass pro-group intentions in some scenarios, this provides ample justification for the theory of participatory intentions beyond the team-reasoning account of cooperation. After all, whereas team reasoning surpasses pro-group intentions with regard to guaranteeing successful cooperation only in *some* scenarios, participatory intentions surpass both in *all* scenarios. The theory of participatory intentions therefore best explains and predicts cooperation.

Alternatively, since I focused solely on the third stage of we-mode reasoning,[55] one could view my study as attempting to understand the logical form of the we-intentions resulting from team reasoning. The logical machinery helps to address this question more precisely than what has been done before. I revised the theory by showing that we-mode reasoning at the third stage may be improved by taking these resulting we-intentions to be participatory intentions rather than team-directed intentions.

4.6 Discussion

What is the relation between individual and collective blameworthiness from the perspective of my theory of participatory intentions? My theory of participat-

[55] These three stages have been discussed in § 4.2.2 (see Hakli et al., 2010, § 4).

ory intentions supports the claim that collective intentional action requires that every member adopts the participatory intention. As noted before, I take it that backward-looking responsibility supports the claim that not fulfilling an obligation is a necessary condition for being blameworthy – both for individuals and for groups. From the perspective of intentionality, there are three types of failures that may ground the group's failure to fulfil its collective obligation: (i) *causal failure*: the combination of individual actions did not amount to a group action that fulfils the group's collective obligation; (ii) *absence of the intention*: the group did not collectively intend to fulfil its collective obligation; and finally (iii) *intention to refrain*: the group collectively intentionally refrained from fulfilling its collective obligation.

Let me start with the first failure. I take it that collective blameworthiness entails that the group failed to fulfil an obligation. If we construe the group's collective obligation as performing a group act that promotes deontic ideality, then this means that they failed to perform a deontically optimal group act. As noted in Chapter 2, the fulfilment of a collective obligation is neither sufficient nor necessary for the fulfilment of individual obligations (§ 2.2). If a group is able to communicate and agree upon a group plan in order to fulfil its collective obligation, it should adopt a *good plan*, which consists only of deontically optimal group acts and is interchangeable (§ 2.4). If a good plan has been adopted, the group's failure to fulfil its collective obligation logically entails that at least one member failed to fulfil her member obligation, and, moreover, it entails that at least one member failed to fulfil her individual obligation in the updated decision context.

My discussion of intentionality and several we-intentions sheds some additional light on this case. Agreement-based joint action is one of the paradigm cases of collective intention (see § 2.3). If a group plan has been adopted, then we may say that the group collectively intends to perform one of the group actions in

the plan. Recall that an individual agent is required to promote the realization of her individual intention. More specifically, under these circumstances, this means that if an agent adopts the team-directed intention, then she is committed to performing an individual action that is her component of one of the group actions in the plan. Accordingly, an agent acts in a way that is faithful to the team-directed intention *if and only if* she fulfils her member obligation. In combination with our insights noted in Chapter 2, this entails that, if the group adopted a good plan, failing to fulfil its collective obligation logically entails that at least one of the members failed to act according to her team-directed intention.

I showed that participatory intentions refine the action recommendations of team-directed intentions, that is, an agent acts according to the participatory intention only if she acts in a way that is faithful to the team-directed intention. Hence, when a group regulates its joint action by adopting a group plan, an agent acts according to the participatory intention *only if* she fulfils her member obligation. Or, conversely, if an agent fails to fulfil her member obligation then she does not act in a way that is faithful to the participatory intention.[56] In sum, this means that, if the group adopted a good plan, failing to fulfil its collective obligation logically entails that at least one of its members failed to act according to the participatory intention.

Therefore there are two important ways in which an individual may have failed to fulfil her member obligation. First, it may be that a member did not adopt the participatory intention. One could say that the fault is in the individual agent's intention. It is crucial to note that this is different from intending not to participate, which is a purposeful act aimed at not participating. For instance, the former, not the latter, includes the possibility that the member is wholly unaware of their joint endeavour. Second, it may be that the agent adopted the participatory

[56] The entailments mentioned so far even hold for group plans that fail to be interchangeable.

intention yet somehow failed to act accordingly. One could say that the failure is in the execution. Roughly stated, the member failed to carry out her part of jointly fulfilling the collective obligation even though she intended to play her part.

Be that as it may, in cases where the group cannot communicate or agree upon a good plan to fulfil its collective obligation, the relation between fulfilling its collective obligation and fulfilling a member obligation may be obscured. Consequently, in such cases, the connection between fulfilling its collective obligation and acting according to the participatory intentions may be unclear.

Nonetheless, I think it is important to note that participatory intentions enhance member obligations, with respect to successfully fulfilling its collective obligation. Recall that participatory intentions refine the action recommendations of team-directed intentions. I argued that an agent fulfils her member obligation if and only if she acts according to the team-directed intention. If it is impossible for the group to adopt a good plan, the group may fail to fulfil its collective obligation even though each member acts according to her team-directed intention. However, participatory intentions surpass team-directed intentions in selecting cooperatively rational outcomes (Result 3). After all, in some scenarios, participatory intentions guarantee that the collective obligation is fulfilled whereas team-directed intentions fail to do so. Accordingly, in some cases, acting according to participatory intentions may guarantee successfully fulfilling a collective obligation, whereas fulfilling member obligations fails to do so.

The second type of failure is the absence of an intention: the group might be collectively blameworthy for *not collectively intending* to fulfil its collective obligation. In this chapter, I only argued for one particular direction of the relation between collective and individual intentions, viz. a group collectively intends to φ only if each member adopts the corresponding participatory intention. The other direction of the relation, however, also seems plausible. Christopher Kutz, for instance, writes:

> So long as the members of a group overlap in the conception of the collective end to which they intentionally contribute, they act collectively, or jointly intentionally. I call this the minimalist conception of joint action. (Kutz, 2000, p. 90)

Overlapping intentions form the basis for Kutz's analysis of collective intentional action. That is, Kutz argues, whenever a set of individuals all adopt the participatory intention, then they act jointly intentionally. Or, conversely, a group does not act collectively intentionally *only if* some of its members did not intentionally participate. If this is correct, then we can conclude that if a group is collectively blameworthy for not collectively intending to fulfil its collective obligation, then at least some members did not intend to participate in jointly fulfilling its collective obligation.

It is unclear whether the absence of a member's participatory intention prompts individual blameworthiness. However, one might want to argue that, in certain cases, this constitutes culpable recklessness on the part of the member. That is, following Dubber (2002) on the mode of acting recklessly, one might argue that a law-abiding member would have intentionally participated in jointly fulfilling a collective obligation. The study of the scope and application of this suggestion has to be left for future research.

Finally, the third failure consists in the group *intentionally failing to fulfil its collective obligation*. In that case, my theory of participatory intentions entails that each member intentionally participates. That is, every member intended to participate in this collective failure. To address these cases, Christopher Kutz writes:

> The Complicity Principle assumes a different view of collective action. Intuitively, marginally effective participants in a collective harm are accountable for the victims' suffering, not because of the individual differences they make, but because their intentional participation in

> a collective endeavor directly links them to the consequences of that endeavor. The notion of participation rather than causation is at the heart of both complicity and collective action. (Kutz, 2000, p. 138)[57]

Intentionally participating in a collective endeavour may therefore induce *inclusive authorship* for the consequences brought about by that endeavour.[58] It is important to contrast this sense of authorship with a mere causal sense of co-authorship.[59] Following Kutz, it could be argued that each member can be held responsible for her intentional participation in a collective harm. Therefore, every member would be *complicit* in the group's harmful group act. The study of the application and extent of these considerations of inclusive authorship are beyond the scope of the current chapter. Nonetheless, these arguments highlight how my theory could be used to relate collective blameworthiness to members' complicity.

[57] Tracy Isaacs (2011, p. 102) does not build on participatory *intentions*; rather, she writes: "The action of an individual may warrant descriptions that invoke collective content. In some situations, this collective content may, as in the individual case, make reference to act descriptions that guide us to moral judgments. . . . An individual act of murder, performed as a part of a larger initiative of genocide, has a distinct moral character that it would lack outside the genocidal context because in genocide it is performed with the aim of destroying a group."

[58] Compare Tomasello (2016, p. 107): "claims of inclusive authorship are licensed by the fact that each member's acts are explained by the overlapping goal of realizing the group's plans".

[59] Braham and van Hees (2012) base their account of moral responsibility on a purely causal sense of co-authorship. Their account includes a condition of "Causal Relevancy Condition", which "is necessary because we cannot say that an outcome bears the stamp of authorship of a person if the person played no causal role in bringing it about" (Braham and van Hees, 2012, pp. 605–606). In § 3.6 I highlighted how causal co-authorship may be used to ground individual responsibility by concluding that, from the perspective of my theory of collective know-how, collective blameworthiness entails "that at least one member knowingly risked that she is causally responsible for the group's failure to fulfil its subjective collective obligation".

Appendix D

Joint Action, Participatory Intentions, and Team Reasoning

D.1 Modal Logic of Agency and Intentionality: Proofs

Lemma D.1. *Let $p \in \mathcal{P}$, and let $\mathcal{H}$ be a group of agents. Then $[\mathcal{H}\ \mathsf{prom}]p$ cannot be characterized by a formula ψ in $\mathfrak{L}_{STIT}$. The incorporation of the $[\mathcal{H}\ \mathsf{prom}]$-operator in $\mathfrak{L}_{ISTIT}$ entails that it is a proper extension of $\mathfrak{L}_{STIT}$.*

Proof. In modal logic, the standard technique for proving non-characterizability is to use bisimulations. It therefore seems better to use standard modal-logical models, viz. STIT models, rather than game models to prove this result. Given the correspondence between game models and STIT models, the result transfers to game models (see §§ 1.2.2–1.3).

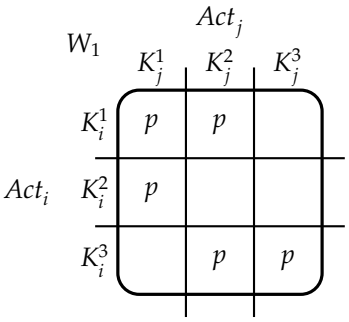
To prove this lemma, I focus on $[i\ \mathsf{prom}]p$ and use the two STIT models depicted in Figure D.1. The bisimulation $Z \subseteq W_1 \times W_2$ is given by the following: for any $w \in W_1$ and any $v \in W_2$ it holds that

wZv iff exactly the same propositions hold at w and v.

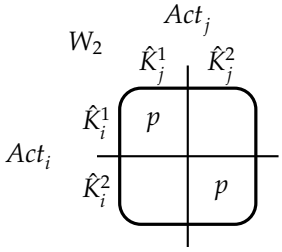
It is now routine to check that Z is indeed a bisimulation (with respect to the language $\mathfrak{L}_{STIT}$). Let us prove one property:

(∗) $w_1 Z v_1$ and $w_1 \in Act_i(w_2)$ implies that there is a $v_2 \in W_2$ such that $v_1 \in Act_i(v_2)$ and $w_2 Z v_2$.

Since, in both $\mathcal{M}_1$ and $\mathcal{M}_2$, every row contains both a p-world and a $\neg p$-world. This property immediately follows.



(a) *The initial model* $\mathcal{M}_1$.



(b) *The constructed model* $\mathcal{M}_2$.

Figure D.1

Finally, I argue by contradiction to prove the lemma. Suppose that $\psi \in \mathfrak{L}_{STIT}$ characterizes $[i \ \mathsf{prom}]p$. Note that $[i \ \mathsf{prom}]p$ holds in any world in $\mathcal{M}_2$. Hence, by assumption, ψ holds at any world in $\mathcal{M}_2$. Take any $w \in K_i^2 \cap K_j^2$. Remark that $\mathcal{M}_1, w \vDash \psi$ since Z is a bisimulation (with respect to $\mathfrak{L}_{STIT}$). However, note that $[i \ \mathsf{prom}]p$ does not hold at $\mathcal{M}_1, w$. Contradiction. We conclude that $[i \ \mathsf{prom}]p$ cannot be characterized by an $\mathfrak{L}_{STIT}$-formula.

□

Lemma D.2. *Let S be a game model, let $\mathcal{H}$ be a group of agents, and let $\varphi \in \mathfrak{L}_{ISTIT}$. Then the following are equivalent:*

1. *there is an $a_{\mathcal{H}} \in A_{\mathcal{H}}$ such that for every $b_{\mathcal{H}} \in A_{\mathcal{H}}$ it holds that $a_{\mathcal{H}} \succeq_{\varphi} b_{\mathcal{H}}$ (i.e. $A_{\mathcal{H}}$ has a $\succeq_{\varphi}$-*maximal *element);*

2. *in* S, $[\mathcal{H}\ \mathsf{prom}]\varphi$ *is characterized by* $[\mathcal{H}\ \mathsf{stit}](\langle -\mathcal{H}\ \mathsf{stit}\rangle\varphi \to \varphi)$.

Proof. Note that the following are equivalent: (i) $S, a \vDash [\mathcal{H}\ \mathsf{stit}](\langle -\mathcal{H}\ \mathsf{stit}\rangle\varphi \to \varphi)$, (ii) for every $c_{-\mathcal{H}}$ it holds that if $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \langle -\mathcal{H}\ \mathsf{stit}\rangle\varphi$ then $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$, (iii) for every $c_{-\mathcal{H}}$ it holds that if there is a $d_{\mathcal{H}} \in A_{\mathcal{H}}$ such that $S, (d_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$ then $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$. It follows that if $S, a \vDash [\mathcal{H}\ \mathsf{stit}](\langle -\mathcal{H}\ \mathsf{stit}\rangle\varphi \to \varphi)$ then for every $b_{\mathcal{H}} \in A_{\mathcal{H}}$ it holds that $a_{\mathcal{H}} \succeq_{\varphi} b_{\mathcal{H}}$. The converse is also easy to see. Suppose that $a_{\mathcal{H}}$ is such that for every $b_{\mathcal{H}} \in A_{\mathcal{H}}$ it holds that $a_{\mathcal{H}} \succeq_{\varphi} b_{\mathcal{H}}$. If $c_{-\mathcal{H}}$ and $d_{\mathcal{H}}$ satisfy $S, (d_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$, then $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$ holds because $a_{\mathcal{H}} \succeq_{\varphi} d_{\mathcal{H}}$.

(1. $\Rightarrow$ 2.) Assume 1. Since there is a $\succeq_{\varphi}$-maximal element in $A_{\mathcal{H}}$, it is the case that, for every $a \in A$, $S, a \vDash [\mathcal{H}\ \mathsf{prom}]\varphi$ holds if and only if for every $b_{\mathcal{H}} \in A_{\mathcal{H}}$ it holds that $a_{\mathcal{H}} \succeq_{\varphi} b_{\mathcal{H}}$. (This equivalence does not hold if there is no maximal element.) Hence, $S, a \vDash [\mathcal{H}\ \mathsf{prom}]\varphi$ if and only if $S, a \vDash [\mathcal{H}\ \mathsf{stit}](\langle -\mathcal{H}\ \mathsf{stit}\rangle\varphi \to \varphi)$.

(2. $\Rightarrow$ 1.) Assume 2. First note that there is an admissible group action, say $a_{\mathcal{H}}$. Then $S, a \vDash [\mathcal{H}\ \mathsf{prom}]\varphi$ and, by assumption, $S, a \vDash [\mathcal{H}\ \mathsf{stit}](\langle -\mathcal{H}\ \mathsf{stit}\rangle\varphi \to \varphi)$. Our initial observation entails that $a_{\mathcal{H}}$ is a $\succeq_{\varphi}$-maximal element in $A_{\mathcal{H}}$. □

Proposition 4.1. *Let* φ *be an* $\mathfrak{L}_{ISTIT}$*-formula and let* $\mathcal{H}$ *be a group of agents, possibly a singleton. Then*

1. $\vDash \Diamond[\mathcal{H}\ \mathsf{prom}]\varphi$*, one is always able to promote* φ*,*
2. $\nvDash [\mathcal{H}\ \mathsf{prom}]\varphi \to [\mathcal{H}\ \mathsf{stit}]\varphi$*, promoting* φ *does not entail ensuring* φ*,*
3. $\vDash [\mathcal{H}\ \mathsf{stit}]\varphi \to [\mathcal{H}\ \mathsf{prom}]\varphi$*, guaranteeing* φ *entails promoting* φ*,*
4. $\vDash [\mathcal{H}\ \mathsf{prom}]\varphi \wedge \Diamond\varphi \to \langle \mathcal{H}\ \mathsf{stit}\rangle\varphi$*, promoting* φ *while* φ *is possible entails allowing* φ*,*
5. $\vDash [\mathcal{H}\ \mathsf{prom}]\varphi \wedge \Diamond[\mathcal{H}\ \mathsf{stit}]\varphi \to [\mathcal{H}\ \mathsf{stit}]\varphi$*, promoting* φ *while being able to ensure* φ *entails guaranteeing* φ*,*
6. $\vDash \Diamond[\mathcal{H}\ \mathsf{stit}]\varphi \to ([\mathcal{H}\ \mathsf{prom}]\varphi \leftrightarrow [\mathcal{H}\ \mathsf{stit}]\varphi)$*, if one is able to guarantee* φ*, then promoting* φ *is equivalent to ensuring* φ*.*

Proof. Let $S = \langle N, (A_i), (Int_{\mathcal{H}}), \pi \rangle$ be any game model.

1. Follows from the fact that we only consider finite game forms (see Lemma B.1 in Appendix B).

2. The previous item shows that if this were a validity, then $\vDash \Diamond[i \text{ stit}]\varphi$ would follow. In other words, one is always able to ensure φ, no matter its logical form. This is, however, not always the case.

3. Suppose $S, a \vDash [\mathcal{H} \text{ stit}]\varphi$. Then for any $c_{-\mathcal{H}} \in A_{-\mathcal{H}}$ we have $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$. Hence, for any $b \in A$ and any $c_{-\mathcal{H}} \in A_{-\mathcal{H}}$ it holds that $S, (b_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$ implies $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$. In other words, $a_{\mathcal{H}} \succeq_{\varphi} b_{\mathcal{H}}$, so $a_{\mathcal{H}}$ is admissible with respect to φ.

4. Suppose $S, a \vDash [\mathcal{H} \text{ prom}]\varphi \wedge \Diamond\varphi$. Then there is a profile, say b, such that $S, b \vDash \varphi$. We argue by contradiction that $S, a \vDash \langle \mathcal{H} \text{ stit} \rangle \varphi$: suppose $S, a \nvDash \langle \mathcal{H} \text{ stit} \rangle \varphi$. Then for any $c_{-\mathcal{H}} \in A_{-\mathcal{H}}$ it is the case that $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \nvDash \varphi$. Hence, for any $c_{-\mathcal{H}} \in A_{-\mathcal{H}}$ it holds that $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$ implies $S, (b_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$, vacuously, i.e. $b_{\mathcal{H}} \succeq_{\varphi} a_{\mathcal{H}}$. Moreover, since $S, b \vDash \varphi$, $b_{\mathcal{H}} \not\preceq_{\varphi} a_{\mathcal{H}}$ and therefore $b_{\mathcal{H}} \succ_{\varphi} a_{\mathcal{H}}$. This shows that $a_{\mathcal{H}}$ is not admissible with respect to φ, which contradicts with our assumption that $S, a \vDash [\mathcal{H} \text{ prom}]\varphi$.

5. We argue by contradiction. Suppose $S, a \vDash [\mathcal{H} \text{ prom}]\varphi$ and $S, b \vDash [\mathcal{H} \text{ stit}]\varphi$, yet $S, a \nvDash [\mathcal{H} \text{ stit}]\varphi$. Recall from the proof of item 3 that $S, b \vDash [\mathcal{H} \text{ stit}]\varphi$ entails that $b_{\mathcal{H}}$ weakly dominates every group action. In particular, $a_{\mathcal{H}} \preceq_{\varphi} b_{\mathcal{H}}$. The assumption that $S, a \nvDash [\mathcal{H} \text{ stit}]\varphi$ entails that there is a $c_{-\mathcal{H}} \in A_{-\mathcal{H}}$ satisfying $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \neg\varphi$. In particular, $S, (b_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$ implies that $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$ does not hold. Hence $a_{\mathcal{H}} \prec_{\varphi} b_{\mathcal{H}}$ holds, which contradicts the assumption that $S, a \vDash [\mathcal{H} \text{ prom}]\varphi$ holds.

6. This follows immediately from the previous item and item 3. □

D.2 We-intentions: Proofs

Observation D.1. *Let S be a game model, let $a \in A$ be a profile, let $\mathcal{G}$ be a group of agents, let $i \in \mathcal{G}$, and let $\varphi \in \mathfrak{L}_{ISTIT}$. Then the following are equivalent:*

1. *At a, agent i performs an individual action that is admissible with respect to her team-directed intention. In other words, a_i is admissible with respect to $\langle i \text{ stit}\rangle[\mathcal{G} \text{ prom}]\varphi$ or, equivalently, $S, a \vDash [i \text{ prom}]\langle i \text{ stit}\rangle[\mathcal{G} \text{ prom}]\varphi$ holds.*

2. *At a, agent i performs an individual action that is her component in an admissible group action. In other words, there is a $b_{\mathcal{G}} \in Admissible_S(\mathcal{G}, \varphi)$ such that $a_i = b_i$ or, equivalently, $S, a \vDash \langle i \text{ stit}\rangle[\mathcal{G} \text{ prom}]\varphi$.*

Proof. Note that $\vDash \langle i \text{ stit}\rangle\chi \leftrightarrow [i \text{ stit}]\langle i \text{ stit}\rangle\chi$, since $[i \text{ stit}]$ is an S5-operator. Given item **1** of Proposition 4.1, it holds that $\vDash \Diamond[\mathcal{H} \text{ prom}]\varphi$. Since $\vDash \Diamond\chi \rightarrow \Diamond\langle i \text{ stit}\rangle\chi$ holds, item **6** of Proposition 4.1 entails that $\vDash [i \text{ prom}]\langle i \text{ stit}\rangle[\mathcal{H} \text{ prom}]\varphi \leftrightarrow \langle i \text{ stit}\rangle[\mathcal{H} \text{ prom}]\varphi$ holds, as desired.

□

Result 1. *Let $S = \langle N, (A_i), (Int_{\mathcal{H}}), \pi\rangle$ be a game model. Suppose group $\mathcal{G}$ collectively intends to φ. Let $a \in A$. Then team reasoning admits the individual action a_i if and only if $S, a \vDash \langle i \text{ stit}\rangle[\mathcal{G} \text{ prom}]\varphi$, which is, in turn, equivalent to saying that the individual action a_i is admitted by the team-directed intention.*

Proof. First note that for any profile $a \in A$ it holds that $S, a \vDash [i \text{ prom}]\langle i \text{ stit}\rangle [\mathcal{G} \text{ prom}]\varphi$ if and only if $S, a \vDash \langle i \text{ stit}\rangle[\mathcal{G} \text{ prom}]\varphi$ (by Observation D.1).

Second, recall that team reasoning implies that individual agent *i* first identifies a best combination that the group members can perform, say $a'_{\mathcal{G}}$, and then decides to perform the individual action that is her part in that combination, which is a'_i. Hence, in the context of a collective intention to φ, for any $a \in A$, the following are equivalent: (1) team reasoning admits a_i, (2) there is an $a'_{\mathcal{G}-i} \in A_{\mathcal{G}-i}$ such that $(a_i, a'_{\mathcal{G}-i})$ is a best group action that the group members can perform, (3) there is a $a'_{-i} \in A_{-i}$ such that $S, (a_i, a'_{-i}) \vDash [\mathcal{G} \text{ prom}]\varphi$, and finally (4) $S, a \vDash \langle i \text{ stit}\rangle[\mathcal{G} \text{ prom}]\varphi$.

□

Observation D.2. *Let $\mathcal{G}$ be a group of agents, let $i \in \mathcal{G}$, and let $\varphi \in \mathfrak{L}_{ISTIT}$. Then*

$$\vDash [i\ \mathsf{prom}][\mathcal{G}\ \mathsf{prom}]\varphi \rightarrow [i\ \mathsf{prom}]\langle i\ \mathsf{stit}\rangle[\mathcal{G}\ \mathsf{prom}]\varphi$$

Proof. Observation D.1 shows that $\langle i\ \mathsf{stit}\rangle[\mathcal{G}\ \mathsf{prom}]\varphi$ entails $[i\ \mathsf{prom}]\langle i\ \mathsf{stit}\rangle [\mathcal{G}\ \mathsf{prom}]\varphi$. Items **1** and **4** of Proposition 4.1 immediately show that the former is entailed by $[i\ \mathsf{prom}][\mathcal{G}\ \mathsf{prom}]\varphi$. Hence, the statement follows by transitivity of the implication. □

Result 4. *Participatory intentions surpass pro-group intentions with regard to guaranteeing successful cooperation.*

To prove this result, I rely on the following lemma:

Lemma D.3. *Let S be a game model. Let $a \in A$ satisfy $S, a \vDash \bigwedge_{j\in\mathcal{G}}[j\ \mathsf{prom}][\mathcal{G}\ \mathsf{prom}]\varphi$ and $S, a \nvDash [i\ \mathsf{prom}]\varphi$. Then there is a profile $b \in A$ satisfying:*

1. $S, (b_i, a_{-i}) \vDash [i\ \mathsf{prom}]\varphi$
2. $b_i \preceq_{[\mathcal{G}\ \mathsf{prom}]\varphi} a_i$
3. $S, (b_i, a_{-i}) \vDash \bigwedge_{j\in\mathcal{G}}[j\ \mathsf{prom}][\mathcal{G}\ \mathsf{prom}]\varphi$

Proof. Assume all the mentioned assumptions. Since $S, a \nvDash [i\ \mathsf{prom}]\varphi$, there is a $b \in A$ satisfying $b_i \succ_\varphi a_i$ and $S, b \vDash [i\ \mathsf{prom}]\varphi$. Hence, for any $c_{-i} \in A_{-i}$ it holds that $S, (a_i, c_{-i}) \vDash \varphi$ implies $S, (b_i, c_{-i}) \vDash \varphi$. This entails that for any $c_{-i} \in A_{-i}$, $(b_i, c_{\mathcal{G}-i}) \succeq_\varphi (a_i, c_{\mathcal{G}-i})$ holds. Hence, for any $c_{-i} \in A_{-i}$ it holds that $S, (a_i, c_{-i}) \vDash [\mathcal{G}\ \mathsf{prom}]\varphi$ implies $S, (b_i, c_{-i}) \vDash [\mathcal{G}\ \mathsf{prom}]\varphi$. In other words, $b_i \succeq_{[\mathcal{G}\ \mathsf{prom}]\varphi} a_i$ holds. Since $S, a \vDash [i\ \mathsf{prom}][\mathcal{G}\ \mathsf{prom}]\varphi$, this entails that $b_i \preceq_{[\mathcal{G}\ \mathsf{prom}]\varphi} a_i$ and $S, (b_i, a_{-i}) \vDash [i\ \mathsf{prom}][\mathcal{G}\ \mathsf{prom}]\varphi$ hold. To prove the third item, we need to show that for every $j \in \mathcal{G} - i$, $S, (b_i, a_{-i}) \vDash [j\ \mathsf{prom}][\mathcal{G}\ \mathsf{prom}]\varphi$ holds. This follows immediately from the fact that the j-th component in (b_i, a_{-i}) equals that in a and the assumption that $S, a \vDash \bigwedge_{j\in\mathcal{G}}[j\ \mathsf{prom}][\mathcal{G}\ \mathsf{prom}]\varphi$. □

Proof of Result 4. Let $S = \langle N, (A_i), (Int_{\mathcal{H}}), \pi \rangle$ be a game model. Suppose $\mathcal{G}$ collectively intends to φ, and suppose that pro-group intentions guarantee that a best group action is performed. Let $a \in A$ satisfy $S, a \vDash \bigwedge_{i \in \mathcal{G}} [i \ \mathsf{prom}][\mathcal{G} \ \mathsf{prom}]\varphi$. Using the previous lemma, we can show, by induction on the size of $\mathcal{F} := \{j \in \mathcal{G} \mid S, a \nvDash [j \ \mathsf{prom}]\varphi\}$, that there is a $b \in A$ satisfying (1) $S, (b_{\mathcal{F}}, a_{-\mathcal{F}}) \vDash \bigwedge_{j \in \mathcal{G}} [j \ \mathsf{prom}]\varphi$ and (2) $b_j \preceq_{[\mathcal{G} \ \mathsf{prom}]\varphi} a_j$ for every $j \in \mathcal{G}$. In light of the assumption that pro-group intentions guarantee that a best group action is performed, (1) implies $S, (b_{\mathcal{F}}, a_{-\mathcal{F}}) \vDash [\mathcal{G} \ \mathsf{prom}]\varphi$. Using (2), this can be shown, by induction on the size of $\mathcal{F}$, to imply $S, a \vDash [\mathcal{G} \ \mathsf{prom}]\varphi$. This shows that participatory intentions guarantee that a best group action is performed. □

This page intentionally contains only this sentence.

5

Practical Reasoning, Cooperation, and Responsibility Voids

> *Let group agents be freed from the burden of being held responsible, and the door will open to abuses: there will be cases where no one is held responsible for actions that are manifestly matters of agential responsibility.*
>
> Philip Pettit (2007, p. 196–197)

5.1 Introduction

Many collective outcomes are the product of the actions of different people. Can any individual be held morally responsible for such outcomes? Can any group be held collectively morally responsible for such outcomes? Do responsibility voids exist? That is, are there situations in which the group is collectively morally responsible for some outcome although no member can be held individually morally responsible for it? In committee decision-making, this yields the question of whether there is a voting procedure and a combination of votes such that none

[†] A revised version of this chapter has been published (see Duijf, forthcoming b).

of the members is an appropriate target of moral criticism, regardless of their involvement in bringing about a certain state of affairs. How should we analyse such cases?

Whether collective moral responsibility distributes to its members is a matter of much debate. For example, collective action problems have been examined before in terms of voting paradoxes (Pettit, 2007), causal contributions and information states of the involved individuals (Braham and van Hees, 2012; see also § 3.6), and the nature of the intentions of the participants (Kutz, 2000; see also § 4.6). Inspired by the team-reasoning account of cooperation (Bacharach, 2006), I adopt a new take on the problem and propose to use the distinction between *cooperation* and *competition* to assess the moral responsibility of individuals in collective action contexts.

Team-reasoning theorists rely on different *modes of reasoning* to contrast cooperation and competition. (See § 4.2.1 for an introduction to the idea of team reasoning.) Simply stated, an individual agent faces a competitive decision problem if and only if she reasons individualistically; she faces a cooperative decision problem if and only if she team reasons. To address typical collective action problems, the team-reasoning paradigm needs to be refined to what I call *participatory reasoning*. The details of these reasoning methods need not worry us at present as they will be explicated in the following sections.

In decision theory, it is common to distinguish between making decisions under certainty, risk, and uncertainty.[1] In line with this distinction, I will show that the outcome of participatory reasoning depends on two types of uncertainty.

[1] Traditionally, decision theorists say that "we are in the realm of decision making under:

1. *Certainty* if each action is known to lead invariably to a specific outcome (the words prospect, stimulus, alternative, etc., are also used).
2. *Risk* if each action leads to one of a set of possible outcomes, each outcome occurring with a known probability.
3. *Uncertainty* if either action or both has as its consequences a set of possible specific outcomes, but where the probabilities of these outcomes are completely unknown or are not even meaningful." (Luce and Raiffa, 1957, p. 13 – emphasis added)

External uncertainty: this is where the collective outcome is influenced by expectations regarding external factors. For example, these external factors may include what your opponents are doing, or they may concern physical properties of the world, such as whether it is currently raining. *Coordination uncertainty*: this is where the collective outcome is influenced by expectations regarding in-group coordination. For instance, it may be that there is no correlation device available to successfully coordinate the members' actions. This distinction will be relevant for assessing the conditions that must be met if responsibility voids are to exist in cooperative decision contexts.

To apply the reasoning-based distinction between cooperation and competition, offered by the team-reasoning literature, to collective action problems, I need to provide a *reasoning-based* framework of moral responsibility, or at least a sketch thereof. Roughly stated, to assess whether an individual is responsible for a certain outcome, I study the practical reasoning that has led to her decision.

Based on my reasoning-based framework of moral responsibility for outcomes, I can assess the possibility of responsibility voids. Based on the distinction between competitive and cooperative decision contexts, my arguments are twofold. First, I argue that competitive decision contexts never yield responsibility voids. In such contexts, there cannot be collective moral responsibility for a certain outcome without there being a morally responsible member. Second, I argue that cooperative decision contexts potentially host responsibility voids. To clarify this point, I will illustrate these responsibility voids for both types of uncertainty: external uncertainty and coordination uncertainty. I shall argue that both types of uncertainty host potential responsibility voids, although the conditions for the existence of such voids differ.

The chapter is set out as follows. After some initial ground-clearing with regard to the literature on voting paradoxes in § 5.2, I will provide a sketch of a reasoning-based framework of moral responsibility using practical-reasoning

schemas in § 5.3. In § 5.4, I discuss the team-reasoning account of cooperation using practical-reasoning schemas, and I argue that team reasoning needs to be refined to participatory reasoning in order to address collective action problems. Finally, in § 5.5, I reconsider the existence of responsibility voids and argue for my two main claims: competitive decision problems are free from responsibility voids; cooperative decision problems potentially host responsibility voids, but the conditions that must be met if such voids are to exist differ depending on the type of uncertainty. Since uncertainty plays an important role in my arguments, I explore some ways to work out the details of team reasoning under uncertainty in Appendix E.

5.2 Some Ground-clearing

A prominent problem in the philosophical literature on collective decision-making is the *discursive dilemma*, which is also known as the doctrinal paradox in legal theory (Kornhauser, 1992; Chapman, 1998; the former introduced the term 'doctrinal paradox'). To illustrate the dilemma, suppose a committee of academics, consisting of M1, M2, and M3, is deciding on whether to award tenure to Mr Borderline. Imagine the university's tenure policy requires excellence in research, service, and teaching. Suppose the committee members are to decide on the tenure by first voting on each of these fields of competence, then aggregating their votes on each of these fields by majority, and finally deriving the collective decision on the tenure in line with the university's rules. If the members vote in accordance with Figure 5.1, then it turns out that they collectively decide to award tenure even though they are unanimously opposed.

According to Philip Pettit, this scenario illustrates that there is a dilemma involving which collective decision procedure to endorse. First, the committee members could adopt the procedure stated in the example, which yields awarding tenure. This is commonly referred to as the premise-driven collective decision

	Research r	Service s	Teaching t	Tenure? $r \wedge s \wedge t$
M1	Yes	Yes	No	No
M2	No	Yes	Yes	No
M3	Yes	No	Yes	No
Group	Yes	Yes	Yes	→Yes / ↓No

Figure 5.1: *The discursive dilemma.*

procedure. Second, the committee members could aggregate their judgements on the conclusion, which would yield refusing tenure. This is commonly referred to as the conclusion-based collective decision procedure. Pettit (2001) connects this dilemma to the literature on deliberative democracy and then presents arguments from republican theory that suggest that a premise-based collective decision-making procedure should be adopted to account for the reasons-responsiveness of the collective. Pettit's perspective of the dilemma is different from the employees' personal purview and is best called an *external perspective*. As designers, we may ask what would be a sensible way of aggregating the individuals' judgements into a collective judgement.[2] However, it is important to note that similar scenarios can be given for *any* collective decision procedure that satisfies some intuitive constraints (List and Pettit, 2002, Theorem 1).[3]

The perspective I adopt in this paper is therefore different. Assume that Mr Borderline actually does not satisfy the conditions for tenure. To study potential responsibility voids, it is vital to ask whether any of the members is an appropriate target of moral criticism for the incorrectly awarded tenure. Instead of asking how we may sensibly aggregate individual judgements, I therefore ask whether, under the given circumstances, any of the committee members can be held responsible for awarding tenure. My perspective may be called an *internal* decision-theoretical

[2]This is typically the subject of social choice theory. Sen, in his Nobel Prize lecture, says that the central question motivating social choice theory is this: "How can it be possible to arrive at cogent aggregative judgments about the society (for example, about 'social welfare,' or 'the public interest,' or 'aggregate poverty'), given the diversity of preferences, concerns, and predicaments of the different individuals *within* the society?" (Sen, 1999, p. 349).

[3]List and Pettit (2002, § 4) discuss some, non-ideal, strategies for dealing with the dilemma.

perspective. It includes the study of how an individual should take a personal decision *given a particular collective decision procedure*. Game theory offers a useful framework for studying individual decision-making in such interdependent decision problems, that is, scenarios in which the collective outcome depends on the interaction of several individuals.[4] (See § 1.1 for an introduction to game theory.)

5.3 Reasoning-based Moral Responsibility

To apply the reasoning-based distinction between cooperation and competition (§ 5.4) to collective action problems, it is instructive to sketch an outline of a reasoning-based framework for moral responsibility. For simplicity's sake, I will focus on blameworthiness for outcomes rather than praiseworthiness. Although I think it is useful and important to further develop this reasoning-based framework, for my present purposes this sketch will suffice to highlight its main features and, roughly, its application.

To make a start with constructing an account of moral responsibility that relies on practical reasoning, I aim to use practical-reasoning schemas. Following Gold and Sugden (2007), various forms of practical reasoning can be characterized by *schemas* of practical reasoning.[5] Such schemas illustrate how premises describing the decision context and describing what the agent seeks to achieve should be used to decide which action should be taken. In logic, a valid rule of inference illustrates how premises should be used to derive conclusions. Analogously, the fundamental idea is that a practical-reasoning schema infers conclusions about what the agent ought to do from premises which include propositions about what the agent is seeking to achieve. For instance, when reasoning individualistically,

[4] Although my perspective can be labelled as game-theoretical, it is important to stress that I plan to go beyond standard rational choice theory by endorsing the team-reasoning account of cooperation (Bacharach, 2006; Sugden, 2000).

[5] To be explicit, Gold and Sugden (2007) present and discuss schemes for individual reasoning, collective reasoning, and several schemes for team reasoning. The participatory-reasoning schema (see § 5.4) is my novel addition to this range of practical-reasoning schemes.

the agent first considers the individual actions available to her, assesses these individual actions in terms of their consequences, and then finds the individual action that best furthers her interests.

Figure 5.2 depicts a reasoning schema that characterizes individually instrumental reasoning.[6] The first premise (I1) expresses the available or eligible individual actions. The two premises (I2) and (I3) state that the available individual actions are evaluated on the basis of their possible consequences. Premise (I4) describes what the agent seeks to achieve in terms of a preference over outcomes. The individual-reasoning schema emphasizes that the personal preferences over outcomes (I4) are lifted to preferences over the available individual actions (I○): since I prefer O1 over O2, I prefer choosing A over choosing B. Finally, this lifted preference delivers the conclusion of practical reasoning: because I must choose one of them, this entails that I should choose A (I●).

(I1) I must choose either A or B
(I2) If I choose A the outcome will be O1
(I3) If I choose B the outcome will be O2
(I4) I prefer O1 over O2

(I○) I prefer choosing A over choosing B
(I●) I should choose A

Figure 5.2: *Individual-reasoning schema: (I●) states the conclusion; (I○) states an intermediate conclusion; and (I1)–(I4) state the premises.*

It may be helpful to note that this reasoning schema can be straightforwardly adopted in decision contexts that involve risk, that is, in cases where "each action leads to one of a set of possible outcomes, each outcome occurring with a known probability" (Luce and Raiffa, 1957). To see this, imagine an individual agent

[6] The 'I' is added to indicate that the statement occurs in the individual-reasoning schema. Similarly, for instance, a 'C' will be added to statements that occur in the collective-reasoning schema (see Figure 5.3).

playing a game of chess against a sophisticated computer. For simplicity's sake, let us imagine that she seeks to win the game. To decide on her next move, the individual agent has to consider the individual actions available to her. Plausibly, only individual actions that abide by the rules of chess are admissible for consideration. She then needs to assess the expected consequences of her eligible individual actions. Say she expects that choosing ♙**e4** has a 75% chance of leading to victory, whereas choosing ♕**d6** only has a 30% chance of doing so. Consequently, assuming she endorses something like expected utility maximization, she prefers choosing ♙**e4** over choosing ♕**d6** because it has the best chance of leading to victory. Finally, this will lead her to conclude that she should choose ♙**e4**.

We can see that the individual-reasoning schema can certainly fail to deliver a determinate solution.[7] That is, if individual reasoning delivers $\mathcal{A}$ as the set of recommended individual actions, then, on the one hand, one may think that each individual action $A \in \mathcal{A}$ is permitted, or admitted, by individual reasoning. On the other hand, one may think that the individual agent is required to choose one of the individual actions in $\mathcal{A}$.[8]

Reasoning schemas highlight that there may be different reasons for holding an individual agent responsible for a certain outcome. As mentioned before, I will focus on moral blameworthiness for outcomes. The intuitive idea is that moral blameworthiness requires or relies on bad intentions or culpable ignorance, or at least on moral faultiness.

To illustrate how the individual-reasoning schema may help develop a reasoning-based framework of responsibility for outcomes, let us consider a case of com-

[7] The example of the chess-player could be amended to prove this point: if the agent expects that two individual actions, say ♙**e4** and ♖**g6**, have the same chance of leading to victory, then individualistic reasoning would yield the recommendation to choose either of these.

[8] I consider this indeterminacy an asset of the individual-reasoning schema. I agree with Bacharach (2006, p. 60, § 8.3) that a theory should only be determinate when our intuitions are. Compare Harsanyi and Selten (1988, p. 13): "Clearly a theory telling us no more than that the outcome can be any one of these equilibrium points will not give us much useful information. We need a theory selecting one equilibrium as the solution of the game."

mittee decision-making. A hiring committee consists of three members, Marie, Malcolm, and Myra, and it has to make a choice about which of three candidates, Alice, Bill, and Charlie, is to join the company board. Committee decisions are made by majority, and with Marie, the chairperson, breaking ties. The committee members vote for Alice, Bill, and Charlie, respectively. The result is that Alice wins. Let us suppose that Alice turns out to be a bad candidate for the company board.

Is Malcolm blameworthy for the outcome that Alice wins? I provide three complementing analyses of this scenario; each emphasizes a particular type of concern and each is associated with some premises of the individual-reasoning schema.[9] First, Malcolm's preferences may have been such that it was irrelevant for him whether Alice wins. It may, for instance, have been the case that he randomly decided on his vote. In this case, Alice's winning played no role in his practical reasoning. It seems implausible to say that Malcolm is morally blameworthy for the result that Alice wins if it played no role in his practical reasoning. To assess Malcolm's responsibility in more detail, we need to ask whether it was reasonable for him to be indifferent about Alice's winning. He would be morally blameworthy for Alice's winning only if it was unreasonable for him to be so indifferent.[10] If it was unreasonable, then this would illustrate a moral failure in premise (I4) of the individual-reasoning schema, which describes what Malcolm seeks to achieve.

Second, Malcolm's preferences may have been such that he sought to let Alice lose. In such a case, Alice's losing played a vital role in his practical reasoning. How could voting for Bill have been permitted by individual reasoning? For example, he could have voted as he did if he expected that Myra would also

[9] Two of these concerns roughly correspond to what Braham and Van Hees (2011) call "normative voids" and "epistemic voids". My first concern has no counterpart in their theory.

[10] Indeed, whether Malcolm actually made a *causal* contribution to Alice's winning is irrelevant. It is only important that it played a role in his practical reasoning. Causality, however, is relevant when he assesses the available individual actions in terms of their consequences.

vote for Bill. In that case, he would expect that his vote would yield Bill's winning. However, his expectation regarding Myra's vote was wrong, leading to Alice's winning. This illustrates a moral failure in premise (I2) or (I3) of the individual-reasoning schema, which describes Malcolm's expectation regarding the consequences of voting for Bill. So we may say that Malcolm is morally blameworthy for Alice's winning: his false expectation led to Alice's winning. To assess Malcolm's responsibility in more detail, we need to ask whether it was reasonable for him to expect that Myra would vote for Bill. Simply stated, his moral responsibility for Alice's winning could be undermined if his expectations were reasonable.

Third, Malcolm may have considered only a subset of the viable options. For instance, he may have thought that casting a vote for Charlie was not an eligible option for him. In this case, Malcolm basically only considered voting for Bill or Alice, then reasoned so as to minimize the expectation that Alice wins, and finally decided to vote for Bill. It then seems plausible to say that Malcolm is not morally blameworthy for Alice's winning. To refine the assessment of moral responsibility, we may ask whether it was reasonable for him to think that voting for Charlie was not an eligible option. Simply stated, he would be morally blameworthy for Alice's winning if he should have known that voting for Charlie was, in fact, an eligible option. The blameworthiness would correspond to a moral failure in premise (I1) of the individual-reasoning schema, which states the eligible individual actions.[11]

[11] It should be clear from the above discussion that my view includes the possibility that the committee members try to 'manipulate' the collective outcome by voting untruthfully. This may contradict one's intuitions. However, for the assessment of moral responsibility it is wholly unclear why truthful votes should take moral priority in interdependent decision problems. I do not wish to claim that every form of strategic decision-making is morally acceptable – only some forms may be. I refer the interested reader to the paper by Dowding and Van Hees (2008) titled "In praise of manipulation".

5.4 Cooperation

To characterize competition and cooperation I rely on the team-reasoning account of cooperation (Bacharach, 2006; Sugden, 2000; Gold and Sugden, 2007). (See § 4.2.1 for an introduction to team reasoning.)[12] The team-reasoning account of cooperation appeals to the *reasoning method* by which an individual agent reasons about what to do.[13] An individual agent engaged in team reasoning "works out the best feasible combination of actions for all the members of her team, then does her part in it" (Bacharach, 2006, p. 121). In this way, team reasoning offers an alternative to standard individualistic reasoning in collective action problems. The core idea of team reasoning is that an individual asks herself 'What should *we* do?' rather than 'What should *I* do?'. Team reasoning hence relies on a we-perspective.[14]

Before we discuss team reasoning, note that the individual-reasoning schema can be straightforwardly amended for groups: the collective-reasoning schema in Figure 5.3 illustrates that the preferences over outcomes are lifted to the available group actions and the schema highlights that it may fail to deliver a determinate solution. The only difference with the individual-reasoning schema is the level of agency: in the individual-reasoning schema, individual preferences over outcomes are lifted to preferences over individual actions; in the collective-reasoning

[12] The current treatment differs from that in § 4.2.1 in that it focuses on the *reasoning methods* by using practical-reasoning schemas.

[13] Bacharach (2006, Ch. 1) and Sugden (2000, §§ 2, 3, 7, and 8) argue that orthodox rational choice theory needs to be augmented with a collectivistic reasoning method to address certain cooperation problems – most notably, the so-called Hi-Lo game. For a brief rehearsal of these arguments see § 4.2.1.

[14] Important precursors within ethical theories of utilitarianism include those of Hodgson (1967) and Regan (1980). Later, Sugden (1991, 1993) fruitfully introduced team reasoning to the field of game theory. Similar ideas have been proposed by Bacharach (1999); Anderson (2001); Hurley (1989); and Gilbert (1989). As discussed in § 4.2.2, Gold and Sugden (2007) and Hakli et al. (2010) have recently connected the team-reasoning literature to theories of collective intentionality .

schema collective preferences over outcomes (C6) are lifted to preferences over group actions (C○). Given the symmetry, the collective-reasoning schema is valid if and only if the individual-reasoning schema is.[15]

(C1)	We must choose (A,A), (A,B), (B,A), or (B,B)
(C2)	If we choose (A,A) the outcome will be O1
(C3)	If we choose (A,B) the outcome will be O2
(C4)	If we choose (B,A) the outcome will be O3
(C5)	If we choose (B,B) the outcome will be O4
(C6)	O1 and O2 best satisfy our collective preferences
(C○)	Choosing (A,A) and choosing (A,B) best satisfy our collective preferences
(C●)	We should choose (A,A) or choose (A,B)

Figure 5.3: *Collective-reasoning schema: (C●) states the conclusion; (C○) states an intermediate conclusion; and (C1)–(C6) denote the premises.*

It is helpful to emphasize that collective reasoning relies on group preferences. In line with the team-reasoning literature (see Bacharach, 1999; Sugden, 2000), I will assume that group preferences are given.[16] As such, my study is independent of any specific account of group preferences.

Proponents of team reasoning have adopted the reasoning schema depicted in Figure 5.4 for group members adopting a we-perspective. It may be helpful to

[15]Gold and Sugden (2007, p. 123 – terminology adapted) write: "We assert that [the individual-reasoning schema and the collective-reasoning schema] are both forms of valid instrumental reasoning and that they are valid for agents, defined as those entities that use these modes of reasoning (individual agents in one case, group agents in the other)."

[16]There has been some discussion on the relation between group preferences and individual preferences. For example, Gold (2012, p. 195) writes that Bacharach "allowed in principle that the group objective might be welfare decreasing for some members", and according to Sugden (2000, p. 176) "the preferences of a team are not necessarily reducible to, or capable of being constructed out of, the preferences that govern the choices that the members of the team make as individuals". Nevertheless, Bacharach (1999, terminology adapted) writes that, given a group of individuals $\mathcal{G}$, "if it happens that [the group preferences are] Paretian, then it is good for everyone in $\mathcal{G}$ if everyone in $\mathcal{G}$ team reasons".

note that there is a subtle difference between team reasoning and team-directed reasoning. For a given group $\mathcal{G}$, Sugden (2000, p. 195 – emphasis added) writes: "Suppose the following two conditions are satisfied. First, each individual $i \in \mathcal{G}$ engages in *team-directed reasoning* with respect to $\mathcal{G}$ and [the group's preferences]. Second, each individual $i \in \mathcal{G}$ has full team confidence with respect to $\mathcal{G}$ and [the group's preferences]. Then ... the team engages in *team reasoning*."[17] Because I will not study the network of common beliefs required for full team confidence, my study can be viewed as focusing on team-directed reasoning. Nonetheless, my study refines a vital component of team reasoning.

(C1)	We must choose (A,A), (A,B), (B,A), or (B,B)
(C2)	If we choose (A,A) the outcome will be O1
(C3)	If we choose (A,B) the outcome will be O2
(C4)	If we choose (B,A) the outcome will be O3
(C5)	If we choose (B,B) the outcome will be O4
(T6)	(A,A) *uniquely* maximizes our collective preferences
(T•)	Each of us should choose her component of (A,A)

Figure 5.4: *Team-reasoning schema: (C1)-(C5) and (T6) state the premises, and (T•) states the conclusion.*

The team-reasoning schema can generally not be applied to typical collective action problems. To illustrate this defect and clarify the problem at this stage, consider the *ambiguous Hi-Lo game* depicted in Figure 5.5. It seems that a satisfactory theory of cooperation should recommend *high* to player 2 and should recommend that player 1 chooses either X or Y, rather than *low*. Team reasoning, unfortunately, fails to deliver either of these recommendations because it relies on the premise that there is a *unique* best group action available, as stated in premise

[17] Note that Sugden (2000, p. 195) writes: "Team-directed reasoning is something that one individual can engage in, independently of any others."

(T6).[18] In collective action problems, it is often the case that there are multiple ways to successfully coordinate the individual actions of the group members. For example, in committee decision-making under the majority rule, it does not matter *who* votes for a particular option, it only matters that enough members vote accordingly. Similarly, in threshold public goods games, all strategy profiles in which sufficiently many individuals contribute is adequate. So team reasoning needs to be refined in order to apply it to collective action problems.

		Player 2: High	Player 2: Low
Player 1	X	3, 3	0, 1
	Y	3, 3	1, 1
	Low	0, 0	2, 2

Figure 5.5: *The ambiguous Hi-Lo game.*

To overcome this defect in team reasoning, I aim to contribute to theories of cooperation by augmenting the team-reasoning paradigm to what I call *participatory reasoning*. Roughly stated, participatory reasoning consists of two stages. First, a participatory reasoner considers the group actions available to them, assesses these group actions in terms of their consequences, and finds the group actions that best further their common or collective interest. Second, a participatory reasoner then considers the individual actions available to her, assesses these individual actions in terms of whether they yield a best group action, and then

[18] Gold and Sugden (2007, p. 125), for example, write: "[T]he schema yields conclusions only when a profile that is the unique maximizer of the team payoff function exists." In addition, Sugden (2000, p. 193) writes: "If two or more different combinations of strategies yield exactly the same utility for the team, this decision rule fails to determine what each individual should do."

chooses an individual action that promotes the realization of a best group action. Accordingly, a member who endorses participatory reasoning can be thought of as aiming to participate in a best group action.[19] If there are several best group actions, participatory reasoning can be viewed as yielding an individual action that maximizes the expectation of realizing a best group action. Figure 5.6 depicts the participatory-reasoning schema, which consists of two stages: the collective level and the individual level. (In Appendix E I discuss how the participatory-reasoning schema can be used to explore ways of specifying team reasoning under uncertainty.) That is, a participant first adopts the group perspective to answer the question 'What should *we* do?' and then asks herself 'What should *I* do?'.

Can the participatory-reasoning schema do without the problematic uniqueness assumption? At the first stage, the collective-reasoning schema is adopted, which could yield a set of group actions. In the ambiguous Hi-Lo game, at the first stage a participant concludes that we should choose (X, *high*) or (Y, *high*). At the second stage, player 1 considers the options available to her and assesses the associated consequences. Since she seeks to realize a best group action, and because choosing X and choosing Y are both compatible with realizing a best group act whereas choosing *low* is not, participatory reasoning yields that she should either choose X or choose Y. This means that the participatory-reasoning schema can do without the problematic uniqueness assumption and can therefore be applied to typical collective action problems.[20]

[19] In Chapter 4, I study the relation between team reasoning and so-called participatory intentions. These participatory intentions naturally relate to the participatory-reasoning schema: for an individual agent, the adoption of "a participatory intention requires her to adopt as her personal objective that a *best group action is performed* " (Duijf, forthcoming a, pp. 20–21). Compare Kutz (2000, p. 81): "Call this way of conceiving of action a participatory intention: an intention to do my part of a collective act, where my part is defined as the task I ought to perform if we are to be successful in realizing a shared goal. This conception of oneself as contributing to a collective, as manifested in one's deliberation and action, is what lies at the heart of collective action generally, from simple coordination to complex cooperation."

[20] It is important to investigate the conditions under which a member should/will team reason rather than reason individualistically. See § 4.2 for a discussion of this issue. For my current purposes, it suffices to assume that team reasoning is eligible and justified in some collective action problems.

(Stage 1: the collective level)

(C1) We must choose (A,A), (A,B), (B,A), or (B,B)

(C2) If we choose (A,A) the outcome will be O1

(C3) If we choose (A,B) the outcome will be O2

(C4) If we choose (B,A) the outcome will be O3

(C5) If we choose (B,B) the outcome will be O4

(C6) O1 and O2 best satisfy our collective preferences

(C○) Choosing (A,A) and choosing (A,B) best satisfy our collective preferences

(C●) We should choose (A,A) or choose (A,B)

(Stage 2: the individual level for agent 1)

(I1) I must choose A or B

(P2) If I choose A the outcome will be that we choose (A,A) or that we choose (A,B)

(P3) If I choose B the outcome will be that we choose (B,A) or that we choose (B,B)

(C●) We should choose (A,A) or choose (A,B)

(P○) Choosing A best promotes the realization of a best group act

(P●) I should choose A

Figure 5.6: *Participatory-reasoning schema: (P●) states the conclusion; (P○), (C○), and (C●) state intermediate conclusions; and the others state the premises.*

Having refined team reasoning to participatory reasoning, let me end this section by addressing its main question: how do these practical-reasoning schemas relate to the distinction between competition and cooperation? To distinguish between the intentions that lie behind cooperative actions and the mutually consistent intentions that lie behind Nash equilibrium behaviour, Natalie Gold and Robert Sugden argue that

> collective intentions are the product of a distinctive mode of practical reasoning, team reasoning, in which agency is attributed to groups. (Gold and Sugden, 2007, p. 137)

Hence, whether a set of individual actions constitutes a collective intentional group action depends on the reasoning that led to the component individual actions. So what makes a set of individual actions a case of jointly intentional action is the form of reasoning endorsed by the members. Simply stated, if members endorse team reasoning, then the resulting pattern is a cooperative act, but if members endorse individual reasoning, then the resulting pattern is a mere aggregate of individual acts.

In cooperative settings, this means that the group members should endorse participatory reasoning rather than individual reasoning. Robert Sugden, for example, concurs:

> A cooperative morality enjoins each individual to do her part in achieving outcomes that are good for all. . . . [T]he individual does not ask whether her own actions, considered in isolation, yield preferred outcomes. (Sugden, 1993, p. 72)

This highlights that morality may even dictate which reasoning method is apt.[21] Roughly stated, competitive contexts are characterized by the fact that individu-

[21] For instance, Regan (1980, p. 11) develops a theory of cooperative utilitarianism, which "can be summed up in the statement that each agent ought to co-operate, with whoever else is co-operating, in the production of the best consequences possible given the behavior of non-co-operators".

alistic reasoning is apt, and cooperative decision contexts are characterized by the fact that participatory reasoning is apt. This distinction will ground my analysis of potential responsibility voids.

5.5 Responsibility Voids Reconsidered

To distribute moral responsibility, my reasoning-based framework for moral responsibility highlights that it is useful to distinguish between two types of contexts: competitive and cooperative contexts. In other words, my framework distinguishes between endorsing individual or participatory reasoning. I will discuss these two cases in turn.

5.5.1 Competitive Decision Contexts

Do competitive decision contexts leave room for responsibility voids? As argued before, competitive decision contexts are characterized by the fact that the involved agents, or at least some of them, endorse the individual-reasoning schema. Stated boldly, I defend the claim that in a world of reasonable, rational individualists there is no void between collective blameworthiness and the individual blameworthiness of its inhabitants. To be sure, there may be tragedy, but there is no void between, on the one hand, collective responsibility and, on the other hand, members' responsibility.

It could be that all committee members in the discursive dilemma reasoned individualistically and that the premises they endorsed were fully justified. Does this leave a responsibility void with regard to the awarded tenure? There are two, jointly exhaustive, cases to distinguish. First, they could be justified in endorsing individualistic reasoning.[22] This means that they reasonably faced a competitive problem. It is uncontroversial that competitiveness undermines group cohesion and group agency. According to most theories of moral responsibility, this would

[22]For my arguments it does not matter what grounds this justification.

imply that the group cannot be attributed moral responsibility.[23] Since this entails that the group is not *collectively* morally blameworthy for awarding tenure, it follows that there is no responsibility void. The result may certainly be a grave tragedy, but not one in which there is some moral residue that should be distributed.

Second, they could be unjustified in endorsing individualistic reasoning. In this case, one could maintain that the group is collectively morally blameworthy for awarding tenure. For instance, the fact that the group forms a committee designed to decide on whether to award tenure may imply that the group is collectively morally blameworthy for the awarded tenure. On my reasoning-based framework of moral responsibility, this means that the committee members' individual blameworthiness should be traced through the adopted reasoning methods. However, collective intentional group acts require that the members reason participatorily rather than individualistically. Hence, the faultiness is the absence of a collective intentional act, whose absence is constituted by the fact that some members reason individualistically. Therefore any member who reasoned individualistically is an appropriate target of moral blame or sanction. That is, the group's collective moral responsibility for the awarded tenure is distributed to those members who reasoned individualistically.

In either case there is no responsibility void: competitive decision contexts are free from responsibility voids.

5.5.2 Cooperative Decision Contexts

Do cooperative decision contexts leave room for responsibility voids? As discussed before, cooperative decision contexts are characterized by the fact that the involved individuals endorse the participatory-reasoning schema. For sim-

[23] There are few exceptions; Chant (2015, §4–5) argues against this position and proposes the "*revised agency thesis*: If moral responsibility attaches to an entity, then that entity must be an agent or a collective constituted by agents".

plicity's sake, I assume that the decision context and the collective preferences are commonly known.[24] To discuss potential responsibility voids it is helpful to distinguish between two different types of uncertainty that the members of the group may jointly face. This distinction is natural if we note that the participatory-reasoning method treats expectations regarding external factors and in-group coordination differently. This difference gives rise to two types of uncertainty:

- *External Uncertainty*. In the first stage of participatory reasoning, group preferences over outcomes are lifted to preferences over group actions using the agent's expectations regarding external factors. For example, these external factors may include what her opponents are doing, or it may concern physical properties of the world, such as whether it is currently raining.

- *Coordination Uncertainty*. In the second stage of participatory reasoning, preferences over individual actions are derived from the previously obtained best group actions and the agent's expectations regarding individuals inside the group.

Both of these types of uncertainty may leave room for responsibility voids. However, the conditions for the existence of responsibility voids differ.

External Uncertainty

To illustrate the characteristics of cases of external uncertainty it may be helpful to construct a simple game form. Imagine that two academics, Ann and Bob, form a committee that needs to decide on whether to award tenure to Mr Borderline. This time, they do so by only judging the candidate's excellence in research; say it is impossible for them to discuss this pre-vote or to adopt a premise-based decision

[24] Relaxing these conditions could reveal new ways in which responsibility voids may obtain. In accordance with my discussion of moral responsibility using the individual-reasoning schema in § 5.3, it may be helpful to say that these conditions entail that I only consider failures in premises (C2)–(C5), or (P2) or (P3), which describe the expected consequences of the available group acts and of the available individual acts, respectively.

procedure. The collective decision is based on unanimity, that is, the collective decision is *X* if and only if both of them voted for *X*. (This includes the possibility of collective indecision.)[25] This case is depicted in Figure 5.7, where Ann and Bob need to vote 'Y' or 'N', and the external factor is modelled as a third player. The group should coordinate on voting 'Y' or on voting 'N', depending on whether the candidate is an excellent researcher. The best group action therefore depends on some external factor.

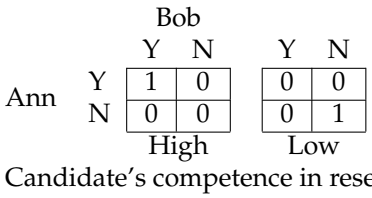


Figure 5.7: *A two-player discursive dilemma, where the members need to decide to award tenure depending on the candidate's excellence in research.*

This case may uphold the possibility of a responsibility void. Such voids are possible if Ann and Bob have diverging beliefs regarding the candidate's excellence in research. Say Ann thinks it is likely that the candidate is an excellent researcher whereas Bob thinks the opposite. Therefore Ann will conclude, in the first stage of participatory reasoning, that they should coordinate on (Y, Y), whereas Bob will conclude that they should coordinate on (N, N). Then, in the second stage, each will reach the opposite conclusion, thereby yielding (Y, N): collective indecision results. Let us suppose that the candidate is an excellent researcher, yet he is not hired because of Ann and Bob's collective indecision.

I will assume that the concept of collective moral responsibility makes sense in this case. Can any of the committee members be held responsible? The outcome results from the members' diverging expectations regarding an external factor, the

[25] Although the story is similar to the discursive dilemma, it should be clear that the structure of the decision context is different.

research excellence of the candidate in this case. This expectation affects the first stage of participatory reasoning in the premises concerning the expected outcomes of the available group acts, viz. (C2)–(C5). One of the committee members must be wrong in his or her expectation, and hence, in the absence of a plausible excuse, the member who has the least accurate expectations is most blameworthy; under the given circumstances, this would be Bob. So there would be no responsibility gap.

However, if Bob has reasonable or justifiable expectations, then he would not be responsible for the collective outcome. His mistake would then be justifiable. The study of these cases is beyond the scope of the current chapter and a discussion of whether certain expectations are reasonable or justifiable has to be left for future work.

Let me briefly clarify how this discussion can be transferred to the original discursive dilemma (Figure 5.1). Suppose the members award tenure in the way stated in the original formulation, but Mr Borderline is actually a poor candidate. This means that the candidate is poor in at least one of the fields of competence. Let's say that the candidate is poor in research. In that case, participatory reasoning helps to clarify that the failure is in the premises (C2)–(C5). For instance, M1 would falsely think that the candidate is excellent in research and that they should decide correspondingly. Accordingly, the committee members who voted for the candidate's excellence in research are most blameworthy; under the given circumstances, those are M1 and M3.

Although this brief discussion might not solve the responsibility distribution in cases of external uncertainty, it is important to note that the literature has largely neglected this type of uncertainty. Moreover, the literature on judgement aggregation (see List and Pettit, 2011, Ch. 2, for a useful overview) seems to

largely focus on cases where the group jointly faces a decision problem under certainty. My discussion conveys the importance of cases of external uncertainty to debates on distributing responsibility.

Coordination Uncertainty

To illustrate the characteristics of cases of coordination uncertainty it may be helpful to construct a simple game form. Imagine that two pedestrians, Clarice and Devin, approach one another in a narrow corridor. They can either keep left or keep right. They are both better off if they both choose the same respective side, otherwise they will have to stop, wasting precious time, or bump into each other. This scenario is depicted in Figure 5.8. Most importantly, this is a scenario in which the pedestrians must coordinate to solve the problem, there are two ways to solve the problem, and they are indifferent about which way they do it.[26] Let us suppose that they fail to coordinate and bump into each other. As noted before, I assume that collective moral responsibility makes sense in this case. Now, let us study whether this case could host responsibility voids, i.e. the absence of individual moral responsibility.

		Devin	
		Left	Right
Clarice	Left	1	0
	Right	0	1

Figure 5.8: *The narrow corridor.*

When endorsing participatory reasoning, the first stage will result in the conclusion that Clarice and Devin should coordinate on either (*left*, *left*) or (*right*, *right*). This underdetermination may result in a responsibility void. Let us see how. Imagine that Clarice expects Devin to walk *left*. At the second stage of participatory reasoning, she will think that her walking *left* is more likely to yield

[26] My personal inspiration for this example comes from Guala (2016, p. 24), who is inspired by the use of game-theoretical models in the philosophical study of conventions by Lewis (1969).

successful coordination, viz. (P2) and (P3). Consequently, she concludes that she should walk *left*. However, if Devin expects Clarice to walk *right*, he will, analogously, conclude that he should walk *right*. Hence, the resulting group action will be (*left*, *right*), which is surely a suboptimal group act. In this case, it is impossible to point out who is at fault. Clarice may respond to Devin in two ways: 'Why did you not *foresee* that I would walk *left*?' or 'Why did you walk *right*?', and hence, one blames the other for a wrong expectation or for a wrong choice. It is, however, unclear who is at fault: when Clarice blames Devin for not expecting her to walk *left*, Devin can react by blaming Clarice for not walking like he expected, that is, for failing to walk *right*. A responsibility void arises.

How can we circumvent such a responsibility void? If communication is possible and agreement is unproblematic, then one way is to align the expectations of group members by letting them agree on a coordinating plan.[27] Raimo Tuomela writes:

> If agreement making is in question, there will also be a publicly existing social (or, if you like, quasi-moral) obligation to participate in joint action. This entailment of an obligation can be regarded as a conceptual truth about the notion of agreement. (Tuomela, 2005, p. 345)

If communication is impossible, an alternative way to address this responsibility void is given by a theory of *salience*. In game theory, the most well-known discussion of salience is given by Thomas Schelling, who writes:

> Most situations provide some clue for coordinating behavior, some focal point for each person's expectation of what the other expects him to expect to be expected to do. (Schelling, 1960, p. 57)[28]

[27] In Chapter 2, I study how a group may regulate its group action by agreeing on a plan. In this case, it is vital to "characterize the conditions under which a group plan successfully coordinates the individual actions of the group members" (Tamminga and Duijf, 2017, p. 187).

[28] Schelling (1960) thinks that these focal points may not be found within a theory of games: "Finding the key, or rather finding a key – any key that is mutually recognized as the key becomes the key

A theory of salience may therefore highlight a particular strategy profile as salient. For example, in the narrow corridor it is plausible to say that in Sweden the salient focal point is that both people walk on their respective right side of the corridor, that is, to coordinate on (*right*, *right*). It may thus be reasonable to expect that both are aware of this focal point. This means that members who have opposing expectations, and therefore aim at coordinating on (*left*, *left*), could be held responsible. The existence of a salient focal point would avoid the type of responsibility void just discussed.

Although agreements and focal points may circumvent this type of responsibility void, it remains unclear whether they are available in *every* case of coordination uncertainty. When these mechanisms are unavailable, my discussion reveals that coordination uncertainty potentially hosts responsibility voids.

5.6 Conclusion

I have discussed the outline of a reasoning-based framework for moral responsibility and highlighted its application in collective action problems. The existence of responsibility voids depends on the nature of the decision context: competitive decision contexts are free of such voids, whereas cooperative decision contexts may host such voids. The conditions for the existence of these voids rest on the type of uncertainty the group faces, that is, either external or coordination uncertainty.

– may depend on imagination more than on logic; it may depend on analogy, precedent, accidental arrangement, symmetry, aesthetic or geometric configuration, casuistic reasoning, and who the parties are and what they know about each other."

This page intentionally contains only this sentence.

Appendix E

Practical Reasoning, Cooperation, and Responsibility Voids

E.1 Team Reasoning under Uncertainty: A Blueprint

In this appendix, I explore the topic of *team reasoning under uncertainty* using the participatory-reasoning schema (see Figure 5.6) as a general guideline. The participatory-reasoning schema is helpful in providing a blueprint for dealing with uncertainty and it is fruitful to briefly discuss some alternative ways of doing so. As such, this appendix is more concerned with raising questions than answering them.

I investigate four dimensions of uncertainty that are important for the idea of team reasoning:

1. Uncertainty regarding the question of whether one should endorse team reasoning rather than individual reasoning.
2. Uncertainty regarding the composition of the group and the question of what the appropriate unit of agency is.
3. External uncertainty.
4. Coordination uncertainty.

To begin the exploration of team reasoning under uncertainty, it is helpful to restate one of the problems for the idea of team reasoning: under which conditions should/will an individual team reason rather than reason individualistically? I do not intend to settle this question; rather, I would like to spell out two ways in which a given theory of the conditions that must be met if one is to endorse team reasoning can be inserted into the reasoning-based framework. First, one could add these conditions to the premises of the participatory-reasoning schema. This is the general approach by Gold and Sugden (2007, schemas 4 and 7, respectively), who discuss several such team-reasoning schemas, which for instance rely on mutual assurance or common knowledge of group identification. Second, one could let participatory reasoning be preceded by a meta-reasoning schema to check whether the relevant conditions are met. That is, one could introduce an extra reasoning schema that is designed for assessing whether the conditions for team reasoning with respect to a group $\mathcal{G}$ are satisfied. One could say that each group that satisfies the relevant conditions is an *admissible unit of agency* and that this meta-reasoning schema is aimed at finding these admissible units of agency.

This leads to the idea that there may be several admissible units of agency. Which one of these admissible units of agency is the appropriate unit of agency for a given individual? How should an individual agent deal with cases where there are several, equally plausible, units of agency? That is, if both group $\mathcal{F}$ and group $\mathcal{G}$ meet the conditions for team reasoning and if an individual agent is a member of both, then should she team reason from the perspective of group $\mathcal{F}$ or from that of group $\mathcal{G}$? At the moment, I do not have a theory that would answer these questions in a satisfactory way, and I do not know of any theory that does so. We may of course demand, by stipulation, that there is a unique unit of agency that satisfies the required conditions.[1]

[1] A condition saying 'it is common knowledge among the members of $\mathcal{G}$ that everyone identifies with $\mathcal{G}$' seems to imply that there is only one group that satisfies the condition (assuming knowledge entails truth and that one can only identify with one entity).

Bacharach (1999) provides a theory of team reasoning that applies to cases where an individual agent may be unsure which team to participate in. He provides a model of so-called "unreliable team interactions", which, roughly stated, add an extra parameter to standard games to model the team an individual identifies with. More specifically, Bacharach (1999, § 2) introduces a probability distribution over tuples consisting of a *participation state* and an outside signal, where participation states range over the groups an individual may identify with.[2] It is important to note that his theory of unreliable team interactions does not solve the previously stated unit-of-agency problem, but it does support a concept of equilibrium given unreliable team identification.

In addition to uncertainty regarding the unit of agency, the participatory-reasoning schema (recall Figure 5.6) is helpful in representing *external uncertainty*. In the first stage, called 'the collective level', a participant considers the group actions available to them (C1), assesses these group actions in terms of their consequences (C2)–(C5), and finds the group actions that best further their common or collective interest (C6)–(C∘). When facing external uncertainty, it is important to note that there are several modes of knowledge that may be inserted in the first stage; I explore five options. First, as I do in this chapter, the participant may use *her own expectations* for reasoning through the collective level, independently of what other team members may think. That is, simply stated, she finds the group action that maximizes the group's utility relative to her own expectations regarding external factors.

Second, she could use the group's *common ground* to assess, for each group action, the prospects of carrying it out. That is, she asks herself what *we* think the possible consequences of *our* group actions are, rather than asking what she

[2] Bacharach (1999, p. 122) writes: "This means that in the present model an agent 'finds herself' participating in a certain team: her participation is not a choice, nor otherwise endogenized."

herself thinks the possible consequences of our group actions are. My theory already relies on group preferences, so one might consider taking the group's common ground as another parameter of participatory reasoning.

Third, computer scientists standardly distinguish between distributed, mutual, and common knowledge.[3] These *modes of group knowledge* may be used to specify the consequences that each available group action may lead to. For example, one could say that a participant needs to assess the available group actions in terms of its commonly known consequences; or in terms of its mutually known consequences; or in terms of its distributively known consequences. Note that, for a given group action K, if the group commonly knows that K excludes a certain outcome, then the group mutually knows that K excludes it. Analogously, if the group mutually knows that K excludes the outcome, then the group distributively knows that K excludes it. So the corresponding sets of possible consequences of the group actions decrease if we move from common knowledge to mutual knowledge to distributive knowledge.

Fourth, if the expectations of the group members are represented by subjective probabilities, one could ask how we may sensibly *aggregate the subjective probabilities* of the members of a group. This aggregated subjective probability may then be used to represent the group's expectations. The aggregation of probability distributions has spawned a large field encompassing economics, psychology, statistics, decision theory, engineering, and risk analysis. Since I have largely ignored subjective probabilities in this work, the reader is referred to a review article by Dietrich and List (2016a).

Fifth, it is important to note that the collective-reasoning schema includes three types of premises – regarding eligible or available group actions, their possible consequences, and a group preference – each of which may be the target of uncertainty. Although uncertainty regarding the consequences and the prefer-

[3] See the textbook treatments of Fagin et al. (2003) and Meyer and van der Hoek (1995) – see also § 3.2.

ences have received some attention in the literature on choice under uncertainty, it seems that uncertainty regarding the available group actions has been largely neglected. What are the available or eligible group actions? For example, in the narrow corridor presented in Figure 5.8, if we assume that it is situated in the Netherlands, then the participants may plausibly think that (*left*, *left*) is not an eligible group action. Therefore, it could be argued that there is no coordination problem from the perspective of the participants: there is a unique *eligible* group action that maximizes the group preferences.

Moreover, we may require that each eligible group action is performable or executable. Under a subjective reading, this means that the set of available group actions consists of those group actions that the group collectively knows how to perform. From the perspective of my theory of collective know-how (see Chapter 3), we could say that 'seeing to it that φ' is an eligible group action if and only if the group collectively knows how to φ. Or, equivalently, since knowing-how relates to action types, we could say that the set of group action types is given by those group actions that the group collectively knows how to perform.

Finally, I briefly reflect on two ways of dealing with *coordination uncertainty*. In the second stage of participatory reasoning, called 'the individual level', a participant assesses the available individual actions in terms of their consequences (P2)–(P3) and finds the individual actions that promote the realization of a best group action (P•). I mention two ways to characterize the possible consequences associated with a particular individual action. First, as I do in this chapter, a participant may use *her own expectations* to represent the possible consequences of her available individual actions, independently of what other group members may think. That is, simply put, she finds the individual action that promotes the realization of a best group action relative to her own expectations.

Second, as noted before, Thomas Schelling writes:

> Most situations provide some clue for coordinating behavior, some focal point for each person's expectation of what the other expects him to expect to be expected to do. (Schelling, 1960, p. 57)

So, alternatively, a participant may ask what *the other expects her to expect to be expected to do*. To illustrate, recall Clarice and Devin's coordination problem (see the discussion of Figure 5.8 in § 5.5.2). This would mean that Clarice asks herself what Devin expects that Clarice expects that Devin expects Clarice to do. One could simplify this and submit that Clarice asks herself what Devin expects Clarice to do. This approach triggers the question 'What is expected of me?' and the associated slogan would be 'Do as expected!'. Although this may initially sound promising, it is crucial to note that this may not solve Clarice and Devin's coordination problem. Suppose Clarice and Devin commonly know each other's expectations. Since Clarice expects Devin to walk *left* and Devin knows this, Devin decides to walk *left*. Analogously, Clarice concludes that she should walk *right*. The resulting group action will therefore be (*right*, *left*), which is surely a suboptimal group act. A responsibility void arises. After all, it is unclear whether any of them is an appropriate target for moral criticism regarding her involvement in bringing about the suboptimal outcome.

Despite deferring a solution to the issue of team reasoning under uncertainty, this finalizes my exploration of the four dimensions of uncertainty that are relevant for the idea of team reasoning.

Conclusion

What is the relation between collective blameworthiness and individual blameworthiness? The simple, though highly uninformative, answer is this: it varies in an intricate way. To curb this complexity, I have built on two central ideas and provided a systematic study of this relation. First, to analyse collective and individual responsibility, I have argued, it is vital to distinguish between member and individual responsibility. After all, what I ought to do as an independent individual may significantly differ from what I ought to do as a member of a particular group. Second, I have relied on the idea that different modes of acting – causally, knowingly, and intentionally – are relevant for levels of culpability. These ideas are summarized in Figure 3 and have appeared throughout the thesis, for example in the distinction between individual obligations and member obligations, knowingly doing simpliciter and knowingly playing your part, individual reasoning and team reasoning, and standard individual intentions and participatory intentions. I will first discuss the contributions of the chapters to answering this main research question and will then discuss several possibilities for future enquiries.

To start, I have shown in **Chapter 2** that the relation between collective blameworthiness and individual blameworthiness is not straightforward. I assumed that blameworthiness entails a failure to fulfil an obligation. It is an important observation that a group could fulfil its collective obligation even though none of its members fulfils her individual obligation. And, conversely, a group may fail

Perspective / *Modality*	**Individual**	**Collective**	**Group member**
Causal	Causal action	Collective causal action	Causally contributory action
Knowingly	Knowingly doing	Collectively knowingly doing	Knowingly contributing
Intentionally	Intentionally doing	Collectively intentionally doing	Intentionally participating

Figure 3: *Modes of acting and different perspectives.*

to fulfil its collective obligation even though each member fulfils her individual obligation. These observations are the inspiration for studying the conditions that must be met if responsibility voids are to exist. That is, cases where the group is collectively blameworthy for a certain outcome even though none of its members is an appropriate target of moral criticism.

We have seen that communication and agreements can help overcome such voids. If agreement is unproblematic, then the group members should adopt a *good* plan to regulate their group action. That is, the group members should adopt an interchangeable and optimal group plan. Such a good group plan guarantees that if each member fulfils her member obligation then the group fulfils its collective obligation. Or, conversely, when a good plan has been adopted, a group fails to fulfil its collective obligation only if at least one of its members fails to fulfil her member obligation. And, moreover, under these conditions, a group fails to fulfil its collective obligation only if at least one of its members fails to fulfil her individual obligation in the updated decision context. Hence, if agreement is unproblematic, there is no responsibility void. So a first condition that has to met if responsibility voids are to exist has been exposed: agreement needs to be problematic or communication needs to be restricted.

In **Chapter 3** I have studied the interplay between action and knowledge. It is important to contrast objective and subjective obligations. Subjective obligations depend on the accessible information: an agent is subjectively obliged to *X* only if she knows how to *X*. The miner's paradox illustrates that subjective obligations and objective obligations may differ. That is, there are cases in which an agent is objectively obliged to *X* while not being subjectively obliged to *X*, and vice versa. Since I assumed that collective blameworthiness relies on a collective obligation, under this subjective reading, collective blameworthiness entails that the group collectively knew how to fulfil its subjective obligation yet did not do it. From the perspective of my theory of collective know-how, under these assumptions, a group fails to fulfil its subjective collective obligation only if at least one of its members knowingly risks that she herself causes the group to fail to fulfil its collective obligation. In other words, she knowingly risks being an author of the collective failure. This yields a *pro tanto* reason for considering that particular member an appropriate target of moral criticism.[1]

In **Chapter 4** I have argued that whenever a group of individuals strives for some joint goal, each group member should adopt a participatory intention. A participatory intention is a standard individual intention with a particular kind of content: an agent who adopts a participatory intention can be taken to be aiming at the realization of a best group action. Given the importance of acting intentionally for levels of culpability, it is vital to distinguish between three ways for intentions to figure in the failure to fulfil the group's collective obligation. The first type of failure is causal faultiness: where the combination of individual actions does not amount to a group action that fulfils the group's collective obligation. As mentioned before, if the group adopted a good group plan to regulate its group action yet fails to fulfil its collective obligation then at

[1] Whether these considerations yield *pro toto* reasons for member blameworthiness is up for debate. I have discussed two reasons that may trump this derivation: epistemic reasons, such as when one knows that some other group members will also not carry out their part, or motivational reasons, such as when one does not want to devalue one's individual interests for the collective good.

least one of its members fails to act in a way that is faithful to her participatory intention. Or, conversely, under these circumstances, if each group member acts in a way that is faithful to her participatory intention then the group succeeds in fulfilling its collective obligation. If agreements are problematic, this logical relation may be obscured. Still, the members of the group have a better shot at fulfilling their collective obligation if each member adopts the participatory intention rather than if each member fulfils her member obligation. In either case, if the adoption of participatory intentions guarantees successful cooperation, it seems plausible to say that the absence of a member's participatory intention induces some level of culpability for that particular member.

The second type of failure is the absence of a collective intention: the group may be collectively blameworthy for not collectively intending to fulfil its collective obligation. If we assume that collective intentions are constituted by overlapping participatory intentions, then the group's collective intention would have been present if each member had adopted the corresponding participatory intention. It is therefore plausible that this type of failure entails that some group members did not intentionally participate in jointly fulfilling the group's collective obligation. Hence, those particular members may be considered appropriate targets of moral criticism for undermining the constitution of the collective intention to fulfil the group's collective obligation.

The third type of failure consists in the group intentionally failing to fulfil its collective obligation. My theory of participatory intentions shows that this entails that each member intentionally participates in this collective intentional omission. Kutz's *Complicity Principle* (2000, p. 122) entails that intentional participation triggers inclusive authorship of the consequences of the collective intentional act. Every member of the group is therefore complicit in the group's harmful group act. This complicity grounds the moral criticism of those members who intentionally participated.

In sum, what these cases show is that the grounds for morally criticizing a member vary. In the first case, the criticism involves causal responsibility: if each group member's conduct had been faithful to the participatory intention then the resulting combination of individual actions is likely to yield a group action that fulfils the collective obligation. In the second case, the ground for moral criticism revolves around the absence of a participatory intention. In the last case, the moral criticism focuses on the intentional participation in a collective wrongdoing and the complicity principle helps establish a connection between the collective and member blameworthiness under these circumstances. The existence of responsibility voids requires that these grounds are insufficient for morally criticizing a member. So the conditions that must be met if responsibility voids are to exist vary across these three types of collective blameworthiness.

In **Chapter 5** I have connected the debate on the existence of responsibility voids to the team-reasoning literature on cooperation. This yielded a disparity: competitive decision contexts are free from responsibility voids, whereas cooperative decision contexts may host them. In addition, the conditions that must be met if responsibility voids are to exist in cooperative decision contexts depend on the type of uncertainty that the group faces: external uncertainty or coordination uncertainty. In cases of external uncertainty, the group member who has the least accurate expectations is most blameworthy. However, if her expectations were reasonable then her blameworthiness may be pre-empted and a responsibility void may result. In cases of coordination uncertainty, theories of agreement-making and salience may help overcome the resulting responsibility void. But if agreements are problematic and a focal point is absent, then cases of coordination uncertainty potentially host responsibility voids.

Having summarized the contributions of the chapters to the overall theme of the thesis, the relation between collective and individual responsibility, I would like to conclude with some applications and extensions for future work.

The present work has focused on the question of whether collective blameworthiness entails that some member is to some extent an appropriate target of moral criticism. It is, however, important to investigate *to what extent* members can be appropriate targets of moral criticism in certain collective action problems. Consider the following example by Matthew Braham and Martin van Hees:

> An example of overdetermination in which we may want to impute different degrees of causal impact to the agents involved is when two firms simultaneously pour toxins into a river with one firm dumping twice as much as the other, but in which the actions of both firms are in itself sufficient to cause a certain harm. (Braham and van Hees, 2009, p. 324)

Braham and Van Hees (2009) provide a theory of degrees of *causation*. In light of the modes of acting that are associated with levels of culpability, it would be interesting to investigate whether a theory of degrees can be given for the modes of *knowingly* and *intentionally* – my work in Chapters 3 and 4 could be taken as a starting point. If so, these theories of degrees of causality, knowingly, and intentionality could be imported into the analysis of individual moral responsibility in collective action contexts.

Degrees of responsibility are especially important for studying moral responsibility – both collective and individual – in *hierarchical groups*. The present work has focused on unstructured groups, but it is plausible that the group's structure plays a role in the distribution of responsibility among the group's members. Consider the following case:

Commanded Killing. Suppose a military commander commands his subordinate to shoot an innocent civilian. The subordinate goes ahead and shoots the

civilian. Can we justify that the commander is to some degree blameworthy? Moreover, can we justify that the commander is blameworthy to a greater extent than the subordinate?[2]

It is intuitive that unstructured groups are more readily open to a uniform distribution of the extent of members' blameworthiness rather than an unequal distribution, which seems more plausible in hierarchical groups. After all, a subordinate acting on the orders of a superior seems less blameworthy than the superior. The central ideas of this thesis may shed some light on the puzzle of moral responsibility in hierarchical groups. For example, with regard to causal responsibility, it may not be clear why the commander is to blame. After all, the commander did not shoot the innocent civilian. However, if we take seriously the idea that collective intentionality may be constituted by the adoption of certain collective decision procedures (List and Pettit, 2011), then it seems that the commander's decision to order the shooting of the innocent civilian constitutes a collective intention to do so. As such, my theory of participatory intentions may signify that both the commander and the subordinate are appropriate targets of moral criticism for their involvement in the shooting. The study of moral responsibility in authority-based structured groups is therefore a viable and proximate extension for future work.

The group's structure could be non-authoritarian and less formal than the *Commanded Killing* example presupposes. Consider the following example:

Opinion Leader. Suppose a committee consisting of three members, Marie, Mel, and Mo, is to decide on a particular proposal by simple majority voting. However, imagine that Mo keeps an eye on Marie and is likely to follow Marie's judgement. Suppose that Marie voted in favour of the proposal, that Mo followed Marie's judgement, and that Mel voted against the proposal. The proposal is therefore

[2]This example is inspired by Himmelreich (2015, Chapter 4, p. 60), who argues that "the commander is an agent of the [shooting]".

accepted by Marie and Mo voting in favour. Suppose it is a bad proposal. Can we justify that Mo is blameworthy? Can we justify that Marie is more blameworthy than Mo?[3]

In future work it will probably be fruitful to relate the study of structured groups to network theory (Newman, 2010). That is, a structured group may be represented by a network of individuals with possible relations between them. The nature of the relation between two individuals may be authoritarian, as in the *Commanded Killing*, or be purposeful influence, as in the *Opinion Leader*. Nonetheless, the application of central concepts from network theory, such as centrality, robustness, and diffusion, may help to address moral responsibility in structured groups. For instance, if a certain outcome is robust in a certain network of individuals then we may say that the individual causal responsibility for the outcome is diminished because each individual is less capable of avoiding the outcome.

A different route for future enquiries would be to investigate, more systematically, what constitutes problematic *disagreement*. As noted in Chapter 2, cases in which agreement is unproblematic are free from responsibility voids, so it is vital to study what constitutes problematic disagreement. To illustrate the complexity of this topic, consider the following example:

> Suppose we have two individuals, Ann and Bob, and a single action of going for a walk that needs to be evaluated with respect to the two relevant (and exhaustive) possibilities, that of it being a hot day and of it being a cool day. Suppose that Ann likes to walk when it's hot, but not when it's cool, and that Bob's preferences are just the reverse. Suppose also that Ann believes that it is most likely to be a hot day and

[3] This example is inspired by the "opinion leader" example of Bovens and Beisbart (2011, § 3.1), who propose a measure of voting power for such cases.

> that Bob believes that it will be cool. Then they may both agree that it is a good idea for them to go walking, even though they disagree about everything that this choice depends on. (Bradley, 2005, p. 223)

This illustrates that there are cases that seem to simultaneously host both disagreement and agreement. After all, Ann and Bob agree it is a good idea for them to go walking, while they disagree about all aspects that this choice depends on. Although the example seems artificial, it should be noted that the underlying problem is relevant for financial markets. Voluntary transactions in financial markets are often based on conflicting beliefs and preferences. After all, in standard rational choice theory, the reason for trading goods is that each participant in the trade expects herself to be better off.

The final direction for future research that I would like to mention is the study of moral responsibility in *institutional contexts*. Institutions have several characteristics. Besides the structural aspects of institutions alluded to before, it is important to note that the stability of institutions is often conceptualized as an equilibrium notion. Francesco Guala and Frank Hindriks, for instance, propose a 'rules-in-equilibrium' conception of institutions; they write:

> Are institutions rules or equilibria of a game? We can now see that the answer is "both": an institution may be considered as an equilibrium or as a rule of the game, depending on the perspective that one takes. (Guala and Hindriks, 2015, p. 185)[4]

To study moral responsibility in an institutional context, it is thus important to investigate equilibria. A group is in equilibrium if its members' beliefs and incentives are stable, that is, if none of its members has an incentive to deviate from equilibrium-play given her beliefs.[5] To illustrate, consider the following problematic equilibrium:

[4] See also Guala (2016).

[5] Note the similarity with the concept of a Nash equilibrium (Definition 1.3). However, this formal notion lacks the network of beliefs that are key in the notion of an equilibrium.

Speeding Drivers. Suppose a number of drivers are all speeding on the motorway. Each driver knows that it would be better if all of them slowed down to the legal speed limit. However, each driver also thinks that if she unilaterally slowed down, then the speed difference would probably result in a severe accident, which may yield several casualties. The group of speeding drivers may therefore be viewed as being in equilibrium.

The example illustrates that it may be risky for each driver to attempt to break out of the established equilibrium. Nevertheless, the group of speeding drivers is in an equilibrium that is suboptimal, which may yield institutional blameworthiness. Can we justify that some or all of the speeding drivers are blameworthy or complicit in this institutional wrongdoing? From the perspective of my theory of collective know-how (Chapter 3), we could say that the group of speeding drivers is blameworthy for speeding while collectively knowing how to achieve the optimal outcome in which each keeps to the legal speed limit. Moreover, in this example we can see that each member knows she is a co-author of the collective failure. The nature of the problem in equilibrium-based collective wrongdoing is epistemic: each member knows that others will not carry out their part in collectively achieving the optimal outcome.

Alternatively, with regard to Chapter 5, it is important to highlight that an equilibrium is standardly viewed as a stable state for *individualistic* reasoners. That is, an equilibrium is a state from which no member has an individual incentive to deviate. However, we have seen that such competitive decision contexts are free from responsibility voids. So either there is no collective blameworthiness or some of the individuals are blameworthy for not endorsing a *we-perspective*, which is highlighted in my reasoning-based analysis of responsibility voids.

One could argue that this case calls for institutional *reform*. That is, the drivers are required to break out of the established equilibrium and move to the optimal solution: each should slow down. If we view an institution as an equilibrium,

then institutional reform requires an update in the behaviour and the expectations of the group members. In some cases, this reform may be achieved by updating the expectations, such as when Sweden authorities announced changing from driving on the left-hand side of the road to the right on 3 September 1967. The emergence of more mundane norms and institutions may, however, rely primarily on opportunities for mutually advantageous behaviour rather than updating expectations. I offer an alternative way to achieve institutional reform: update the group members' perspective from individualistic to community-directed. This would yield a transformation from a competitive context to a cooperative one.

It is to be expected that the relation between collective and individual blameworthiness will not be a straightforward matter. We need to join our research efforts to explore the conditions that lead to the existence of problematic scenarios, such as cases of responsibility voids, in order to remedy potential immoral behaviour or outcomes that may arise in such scenarios or completely avoid them by designing our social interactions accordingly.

This page intentionally contains only this sentence.

Bibliography

Ågotnes, T., V. Goranko, W. Jamroga, and M. Wooldridge (2015). Knowledge and ability. In H. van Ditmarsch, J. Y. Halpern, W. van der Hoek, and B. Kooi (Eds.), *Handbook of Epistemic Logic*, pp. 543–589. London: College Publications.

Ågotnes, T. and Y. N. Wáng (2016). Resolving distributed knowledge. In R. Ramanujam (Ed.), *Proceedings of the Fifteenth Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 37–46.

Alur, R., T. A. Henzinger, and O. Kupferman (2002). Alternating-time temporal logic. *Journal of the ACM 49*(5), 672–713.

Anderson, A. R. (1958). A reduction of deontic logic to alethic modal logic. *Mind 67*(265), 100–103.

Anderson, E. (2001). Unstrapping the straitjacket of 'preference': A comment on Amartya Sen's contributions to philosophy and economics. *Economics & Philosophy 17*(01), 21–38.

Anscombe, G. E. M. (1963). *Intention*. Cambridge: Harvard University Press.

Aumann, R. J. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica 55*(1), 1–18.

Aumann, R. J. (1999). Interactive epistemology I: Knowledge. *International Journal of Game Theory 28*(3), 263–300.

Aumann, R. J. and J. H. Dreze (2008). Rational expectations in games. *The American Economic Review 98*(1), 72–86.

Axelrod, R. M. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Bacharach, M. (1999). Interactive team reasoning: A contribution to the theory of co-operation. *Research in Economics 53*(2), 117–147.

Bacharach, M. (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton: Princeton University Press.

Baltag, A., L. S. Moss, and S. Solecki (1998). The logic of public announcements, common knowledge, and private suspicions. In I. Gilboa (Ed.), *Proceedings of the Seventh Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 43–56. Morgan Kaufmann Publishers.

Bardsley, N., J. Mehta, C. Starmer, and R. Sugden (2010). Explaining focal points: Cognitive hierarchy theory versus team reasoning. *The Economic Journal 120*(543), 40–79.

Bartha, P. (2014). Decisions in branching time. In T. Müller (Ed.), *Nuel Belnap on Indeterminism and Free Action*, pp. 29–56. Springer.

Belnap, N. and M. Perloff (1988). Seeing to it that: A canonical form for agentives. *Theoria 54*(3), 175–199.

Belnap, N., M. Perloff, and M. Xu (2001). *Facing the Future. Agents and Choices in Our Indeterminist World*. Oxford: Oxford University Press.

van Benthem, J. and E. Pacuit (2014). Connecting logics of choice and change. In T. Müller (Ed.), *Nuel Belnap on Indeterminism and Free Action*, pp. 291–314. Springer.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica 52*(4), 1007–1028.

Bicchieri, C. (2006). *The Grammar of Society*. Cambridge: Cambridge University Press.

Blackburn, P., M. De Rijke, and Y. Venema (2001). *Modal Logic*. Cambridge: Cambridge University Press.

Bliss, R. and K. Trogdon (2016). Metaphysical Grounding. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.).

Bolton, G. E. and A. Ockenfels (2000). ERC: A theory of equity, reciprocity, and competition. *The American Economic Review 90*(1), 166–193.

Bovens, L. and C. Beisbart (2011). Measuring voting power for dependent voters through causal models. *Synthese 179*(1), 35–56.

Bradley, R. (2005). Bayesian utilitarianism and probability homogeneity. *Social Choice and Welfare 24*(2), 221–251.

Bradley, R. and M. Drechsler (2014). Types of uncertainty. *Erkenntnis 79*(6), 1225–1248.

Braham, M. and M. van Hees (2009). Degrees of causation. *Erkenntnis 71*(3), 323–344.

Braham, M. and M. van Hees (2011). Responsibility voids. *The Philosophical Quarterly 61*(242), 6–15.

Braham, M. and M. van Hees (2012). An anatomy of moral responsibility. *Mind 121*(483), 601–634.

Brandenburger, A., A. Friedenberg, and H. J. Keisler (2008). Admissibility in games. *Econometrica 76*(2), 307–352.

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

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

Broersen, J. (2009). A stit-logic for extensive form group strategies. In P. Boldi, G. Vizzari, G. Pasi, and R. Baeza-Yates (Eds.), *Proceedings of the International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Volume 3, Washington, pp. 484–487. IEEE Computer Society.

Broersen, J. (2011a). Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of Applied Logic 9*(2), 137–152.

Broersen, J. (2011b). Modeling attempt and action failure in probabilistic stit logic. In T. Walsh (Ed.), *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 792–797. AAAI Press.

Broersen, J. and A. Herzig (2015). Using STIT theory to talk about strategies. In J. van Benthem, S. Ghosh, and R. Verbrugge (Eds.), *Models of Strategic Reasoning*, pp. 137–173. Springer.

Broersen, J., A. Herzig, and N. Troquard (2006). Embedding alternating-time temporal logic in strategic logic of agency. *Journal of Logic and Computation 16*(5), 559–578.

Brown, M. A. (1988). On the logic of ability. *Journal of Philosophical Logic 17*(1), 1–26.

Carr, D. (1979). The logic of knowing how and ability. *Mind 88*(351), 394–409.

Chant, S. R. (2007). Unintentional collective action. *Philosophical Explorations 10*(3), 245–256.

Chant, S. R. (2015). Collective responsibility in a Hollywood standoff. *Thought: A Journal of Philosophy 4*(2), 83–92.

Chapman, B. (1998). More easily done than said: Rules, reasons and rational social choice. *Oxford Journal of Legal Studies 18*(2), 293–329.

Chellas, B. F. (1992). Time and modality in the logic of agency. *Studia Logica 51*(3/4), 485–517.

Ciuni, R. and J. Horty (2014). Stit logics, games, knowledge, and freedom. In A. Baltag and S. Smets (Eds.), *Johan van Benthem on Logic and Information Dynamics*, pp. 631–656. Springer.

Cohen, P. R. and H. J. Levesque (1990). Intention is choice with commitment. *Artificial Intelligence 42*(2), 213–261.

Conradie, W., V. Goranko, and D. Vakarelov (2006). Algorithmic correspondence and completeness in modal logic. I. The core algorithm SQEMA. *Logical Methods in Computer Science 2*(1), 1–26.

Darley, J. M. and B. Latané (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology 8*(4), 377–383.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy 60*(23), 685–700.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Clarendon Press.

Decker, K. S. (1995). *Environment Centered Analysis and Design of Coordination Mechanisms*. PhD Thesis, University of Massachusetss.

Dietrich, F. and C. List (2016a). Probabilistic opinion pooling. In A. Hájek and C. Hitchcock (Eds.), *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

Dietrich, F. and C. List (2016b). Reason-based choice and context-dependence: An explanatory framework. *Economics & Philosophy 32*(02), 175–229.

van Ditmarsch, H., W. van der Hoek, and B. Kooi (2007). *Dynamic Epistemic Logic*. Dordrecht: Springer.

Dowding, K. and M. van Hees (2008). In praise of manipulation. *British Journal of Political Science 38*(1), 1–15.

Dubber, M. D. (2002). *Criminal Law: Model Penal Code*. New York: Foundation Press.

Duijf, H. (2015). Performing conditional strategies in strategic STIT theory. In M. Kaeshammer and P. Schulz (Eds.), *Proceedings of the ESSLLI 2015 Student Session*, pp. 13–24.

Duijf, H. (forthcoming a). Beyond team-directed reasoning: Participatory intentions contribute to a theory of collective agency. *Logique et Analyse*. http://virthost.vub.ac.be/lnaweb/ojs/index.php/LogiqueEtAnalyse/article/view/2120.

Duijf, H. (forthcoming b). Responsibility voids and cooperation. *Philosophy of the Social Sciences*. https://doi.org/10.1177/0048393118767084.

Duijf, H. and J. Broersen (2016). Representing strategies. In A. Lomuscio and M. Y. Vardi (Eds.), *Proceedings of the Fourth International Workshop on Strategic Reasoning*, Volume 218, pp. 15–26. Electronic Proceedings in Theoretical Computer Science.

Duijf, H., J. M. Broersen, and J-J. Ch. Meyer (forthcoming). Conflicting intentions: Rectifying the consistency requirements. *Philosophical Studies*. https://doi.org/10.1007/s11098-018-1049-z.

Elgesem, D. (1993). *Action Theory and Modal Logic*. PhD Thesis, University of Oslo.

Elgesem, D. (1997). The modal logic of agency. *Nordic Journal of Philosophical Logic*.

Epstein, B. (2015). *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. New York: Oxford University Press.

Fagin, R., Y. Moses, M. Y. Vardi, and J. Y. Halpern (2003). *Reasoning about Knowledge*. Cambridge: MIT Press.

Faillo, M., A. Smerilli, and R. Sugden (2013). The roles of level-k and team reasoning in solving coordination games. Cognitive and Experimental Economics Laboratory Working Paper 6-13.

Fantl, J. (2008). Knowing-how and knowing-that. *Philosophy Compass 3*(3), 451–470.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics 114*(3), 817–868.

Feinberg, J. (1970). *Doing and Deserving: Essays in the Theory of Responsibility*. Princeton: Princeton University Press.

Fikes, R. E. and N. J. Nilsson (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence 2*(3), 189–208.

Fodor, J. A. (1968). The appeal to tacit knowledge in psychological explanation. *The Journal of Philosophy 65*(20), 627–640.

Føllesdal, D. and R. Hilpinen (1971). An introduction. In R. Hilpinen (Ed.), *Deontic Logic: Introductory and Systematic Readings*. Dordrecht: D. Reidel Publishing Company.

Gansberg, M. (1964). 37 who saw murder didn't call the police. *The New York Times*.

Gauthier, D. (1975). Coordination. *Dialogue 14*(2), 195–221.

Ghallab, M., D. Nau, and P. Traverso (2004). *Automated Planning: Theory and Practice*. Amsterdam: Elsevier.

Gilbert, M. (1989). *On Social Facts*. London: Routledge.

Gilbert, M. (1990). Walking together: A paradigmatic social phenomenon. *Midwest Studies in Philosophy 15*(1), 1–14.

Gilbert, M. (1999). Obligation and joint commitment. *Utilitas 11*(2), 143–163.

Gilbert, M. (2006a). *A Theory of Political Obligation: Membership, Commitment, and the Bonds of Society*. Oxford: Clarendon Press.

Gilbert, M. (2006b). Who's to blame? Collective moral responsibility and its implications for group members. *Midwest Studies in Philosophy 30*(1), 94–114.

Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies 144*(1), 167–187.

Glick, E. (2011). Two methodologies for evaluating intellectualism. *Philosophy and Phenomenological Research 83*(2), 398–434.

Gold, N. (2012). Team reasoning, framing and cooperation. In S. Okasha and K. Binmore (Eds.), *Evolution and Rationality: Decisions, Co-Operation and Strategic Behaviour*, pp. 185–212. Cambridge: Cambridge University Press.

Gold, N. and R. Sugden (2007). Collective intentions and team agency. *The Journal of Philosophy 104*(3), 109–137.

Goldman, A. I. (1971). The individuation of action. *The Journal of Philosophy 68*(21), 761–774.

Goranko, V., W. Jamroga, and P. Turrini (2013). Strategic games and truly playable effectivity functions. *Autonomous Agents and Multi-Agent Systems 26*(2), 288–314.

Guala, F. (2016). *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton: Princeton University Press.

Guala, F. and F. Hindriks (2015). A unified social ontology. *The Philosophical Quarterly 65*(259), 177–201.

Hakli, R., K. Miller, and R. Tuomela (2010). Two kinds of we-reasoning. *Economics & Philosophy 26*(3), 291–320.

Harsanyi, J. C. and R. Selten (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge: MIT Press.

Hawke, P. (2017). The logic of joint ability in two-player tacit games. *The Review of Symbolic Logic 10*(3), 481–508.

van Hees, M. and O. Roy (2008). Intentions and plans in decision and game theory. In B. Verbeek (Ed.), *Reasons and Intentions*, pp. 207–226. Aldershot: Ashgate.

Herzig, A. and F. Schwarzentruber (2008). Properties of logics of individual and group agency. In C. Areces and R. Goldblatt (Eds.), *The Seventh Conference on Advances in Modal Logic*, pp. 133–149. College Publications.

Herzig, A. and N. Troquard (2006). Knowing how to play: Uniform choices in logics of agency. In P. Stone and G. Weiss (Eds.), *Proceedings of the Fifth International Conference on Autonomous Agents and Multiagent Systems*, pp. 209–216. ACM.

Hilpinen, R. (1971). *Deontic Logic: Introductory and Systematic Readings*. Dordrecht: D. Reidel Publishing Company.

Himmelreich, J. (2015). *Agency as Difference-Making: Causal Foundations of Moral Responsibility*. PhD Thesis, The London School of Economics and Political Science.

Hodgson, D. H. (1967). *Consequences of Utilitarianism*. Oxford: Clarendon Press.

van der Hoek, W. and M. Wooldridge (2003). Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications. *Studia Logica 75*(1), 125–157.

Hollis, M. and R. Sugden (1993). Rationality in action. *Mind 102*(405), 1–35.

Horling, B., V. Lesser, R. Vincent, T. Wagner, A. Raja, S. Zhang, K. Decker, and A. Garvey (1999). The TAEMS white paper. Unpublished manuscript.

Horty, J. F. (1996). Agency and obligation. *Synthese 108*(2), 269–307.

Horty, J. F. (2001). *Agency and Deontic Logic*. New York: Oxford University Press.

Horty, J. F. and N. Belnap (1995). The deliberative stit: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic 24*(6), 583–644.

Horty, J. F. and E. Pacuit (2017). Action types in stit semantics. *The Review of Symbolic Logic 10*(4), 617–637.

Hurley, S. L. (1989). *Natural Reasons*. New York: Oxford University Press.

Isaacs, T. (2011). *Moral Responsibility in Collective Contexts*. New York: Oxford University Press.

Jackson, F. (1987). Group morality. In P. Pettit, R. Sylvan, and J. Norman (Eds.), *Metaphysics and Morality: Essays in Honour of J.J.C. Smart*, pp. 91–110. Oxford: Blackwell.

Jamroga, W. and T. Ågotnes (2007). Constructive knowledge: What agents can achieve under imperfect information. *Journal of Applied Non-Classical Logics 17*(4).

Jamroga, W. and W. van der Hoek (2004). Agents that know how to play. *Fundamenta Informaticae 63*(2-3), 185–220.

Jennings, N. R. (2000). On agent-based software engineering. *Artificial Intelligence 117*(2), 277–296.

Kanger, S. (1971). New foundations for ethical theory. In R. Hilpinen (Ed.), *Deontic Logic: Introductory and Systematic Readings*, pp. 36–58. Dordrecht: D. Reidel Publishing Company.

Karpus, J. and M. Radzvilas (2018). Team reasoning and a measure of mutual advantage in games. *Economics & Philosophy 34*(1), 1–30.

Keynes, J. M. (1921). *A Treatise On Probability*. London: Macmillan & Co.

Kohlberg, E. and J.-F. Mertens (1986). On the strategic stability of equilibria. *Econometrica 54*(5), 1003–1037.

Kolodny, N. and J. MacFarlane (2010). Ifs and oughts. *The Journal of Philosophy 107*(3), 115–143.

Konolige, K. and M. E. Pollack (1993). A representationalist theory of intention. In R. Bajcsy (Ed.), *Proceedings of the Thirteenth International Joint Conference on Artifical Intelligence*, Volume 1, pp. 390–395.

Kooi, B. and A. Tamminga (2008). Moral conflicts between groups of agents. *Journal of Philosophical Logic 37*(1), 1–21.

Kornhauser, L. A. (1992). Modeling collegial courts. II. Legal doctrine. *Journal of Law, Economics and Organization 8*(3), 441–470.

Kutz, C. (2000). *Complicity: Ethics and Law for a Collective Age*. Cambridge: Cambridge University Press.

Lewis, D. K. (1969). *Convention: A Philosophical Study*. Cambridge: Harvard University Press.

Lismont, L. (1993). La connaissance commune en logique modale. *Mathematical Logic Quarterly 39*(1), 115–130.

List, C. and P. Pettit (2002). Aggregating sets of judgments: An impossibility result. *Economics & Philosophy 18*, 89–110.

List, C. and P. Pettit (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Lorini, E., D. Longin, and E. Mayor (2014). A logical analysis of responsibility attribution: Emotions, individuals and collectives. *Journal of Logic and Computation 24*(6), 1313–1339.

Luce, R. D. and H. Raiffa (1957). *Games and Decisions*. New York: John Wiley & Sons.

Markoff, J. (2011). Computer wins on 'Jeopardy!': Trivial, it's not. *The New York Times*.

McCarthy, J. and P. Hayes (1969). Some philosophical problems from the standpoint of artificial intelligence. In D. Michie and B. Meltzer (Eds.), *Machine Intelligence 4*, pp. 463–502. Edinburgh University Press.

Metz, C. (2016). Google's AI wins fifth and final game against Go genius Lee Sedol. *WIRED*.

Metz, C. (2017). A mystery AI just crushed the best human players at poker. *WIRED*.

Meyer, J.-J. C. and W. van der Hoek (1995). *Epistemic Logic for AI and Computer Science*. New York: Cambridge University Press.

Meyer, J.-J. C., W. van der Hoek, and B. van Linder (1999). A logical approach to the dynamics of commitments. *Artificial Intelligence 113*(1), 1–40.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America 36*(1), 48–49.

Nash, J. F. (1951). Non-cooperative games. *Annals of Mathematics 54*, 286–295.

von Neumann, J. and O. Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Newman, M. (2010). *Networks: An Introduction*. New York: Oxford University Press.

Nissani, M. (1997). Ten cheers for interdisciplinarity: The case for interdisciplinary knowledge and research. *The Social Science Journal 34*(2), 201–216.

Okasha, S. (2016). On the interpretation of decision theory. *Economics & Philosophy 32*(3), 409–433.

Osborne, M. J. and A. Rubinstein (1994). *A Course in Game Theory*. Cambridge: MIT Press.

Pacuit, E. and O. Roy (2017). Epistemic foundations of game theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017 ed.).

Parfit, D. (1988). What we together do. Unpublished manuscript.

Payette, G. (2014). Decidability of an xstit logic. *Studia Logica 102*(3), 577–607.

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica 52*(4), 1029–1050.

Perea, A. (2012). *Epistemic Game Theory*. Cambridge: Cambridge University Press.

Perloff, M. and N. Belnap (2011). Future contingents and the battle tomorrow. *The Review of Metaphysics 64*(3), 581–602.

Pettit, P. (2001). Deliberative democracy and the discursive dilemma. *Philosophical Issues 11*, 268–299.

Pettit, P. (2007). Responsibility incorporated. *Ethics 117*(2), 171–201.

Plaza, J. (1989). Logic of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, and Z. W. Ras (Eds.), *Proceedings of the Fourth International Symposium on Methodologies for Intelligent Systems: Poster Session Program*, Charlotte, pp. 201–216. Oak Ridge National Laboratory.

Prior, A. N. (1967). *Past, Present and Future*. Oxford: Clarendon Press.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review 83*(5), 1281–1302.

Rao, A. S. and M. P. Georgeff (1991). Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall (Eds.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, Volume 91, pp. 473–484. Morgan Kaufmann.

Regan, D. (1980). *Utilitarianism and Co-Operation*. New York: Oxford University Press.

Risse, M. (2000). What is rational about Nash equilibria? *Synthese 124*(3), 361–384.

Roy, O. (2009a). A dynamic-epistemic hybrid logic for intentions and information changes in strategic games. *Synthese 171*(2), 291–320.

Roy, O. (2009b). Intentions and interactive transformations of decision problems. *Synthese 169*(2), 335–349.

Russell, S. and P. Norvig (1995). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs: Prentice Hall.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley & Sons.

Savage, L. J. (1972). *The Foundations of Statistics* (2nd ed.). New York: Dover Publications.

Scanlon, T. (1998). *What We Owe to Each Other*. Cambridge: The Belknap Press.

Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge: Harvard University Press.

Schelling, T. C. (2010). Game theory: A practitioner's approach. *Economics & Philosophy 26*(1), 27–46.

Schobbens, P.-Y. (2004). Alternating-time logic with imperfect recall. In W. van der Hoek, A. Lomuscio, E. de Vink, and M. Wooldridge (Eds.), *Logic and Communication in Multi-Agent Systems*, Volume 85 of *Electronic Notes in Computer Science*, pp. 82–93. Elsevier.

Searle, J. (1990). Collective intentions and actions. In P. R. Cohen, J. Morgan, and M. E. Pollack (Eds.), *Intentions in Communication*, pp. 401–415. Cambridge: MIT Press.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory 4*(1), 25–55.

Sen, A. (1999). The possibility of social choice. *The American Economic Review 89*(3), 349–378.

Shoham, Y. (1993). Agent-oriented programming. *Artificial Intelligence 60*(1), 51–92.

Stanley, J. (2011). *Know How*. Oxford: Oxford University Press.

Stanley, J. and T. Williamson (2001). Knowing how. *The Journal of Philosophy 98*(8), 411–444.

Sugden, R. (1991). Rational choice: A survey of contributions from economics and philosophy. *The Economic Journal 101*(407), 751–785.

Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social Philosophy and Policy 10*(1), 69–89.

Sugden, R. (1995). A theory of focal points. *The Economic Journal 105*(430), 533–550.

Sugden, R. (2000). Team preferences. *Economics & Philosophy 16*(2), 175–204.

Sugden, R. (2003). The logic of team reasoning. *Philosophical Explorations 6*(3), 165–181.

Sugden, R. (2010). Opportunity as mutual advantage. *Economics & Philosophy 26*(1), 47–68.

Sugden, R. (2011). Mutual advantage, conventions and team reasoning. *International Review of Economics 58*(1), 9–20.

Sugden, R. (2015). Team reasoning and intentional cooperation for mutual benefit. *Journal of Social Ontology 1*(1), 143–166.

Tamminga, A. (2013). Deontic logic for strategic games. *Erkenntnis 78*(1), 183–200.

Tamminga, A. and H. Duijf (2017). Collective obligations, group plans and individual actions. *Economics & Philosophy 33*(2), 187–214.

Thomason, R. H. (1970). Indeterminist time and truth-value gaps. *Theoria 36*(3), 264–281.

Thomason, R. H. (1984). Combinations of tense and modality. In D. Gabbay and F. Guenthner (Eds.), *Handbook of Philosophical Logic*, Volume 165 of *Synthese Library*, pp. 135–165. Dordrecht: Springer.

Tomasello, M. (2016). *A Natural History of Human Morality*. Cambridge: Harvard University Press.

Tuomela, R. (2000a). Collective and joint intention. *Mind & Society 1*(2), 39–69.

Tuomela, R. (2000b). *Cooperation: A Philosophical Study*. Number 82 in Philosophical Studies Series. Dordrecht: Kluwer Academic Publishers.

Tuomela, R. (2005). We-intentions revisited. *Philosophical Studies 125*(3), 327–369.

Tuomela, R. (2006). Joint intention, we-mode and I-mode. *Midwest Studies in Philosophy 30*(1), 35–58.

Tuomela, R. (2007). *The Philosophy of Sociality*. New York: Oxford University Press.

Tuomela, R. (2013). *Social Ontology: Collective Intentionality and Group Agents*. New York: Oxford University Press.

Turrini, P. (2012). Agreements as norms. In T. Ågotnes, J. M. Broersen, and D. Elgesem (Eds.), *Proceedings of the Eleventh International Conference on Deontic Logic in Computer Science*, pp. 31–45. Springer.

Van De Putte, F., A. Tamminga, and H. Duijf (2017). Doing without nature. In A. Baltag, J. Seligman, and T. Yamada (Eds.), *Proceedings of the International Workshop on Logic, Rationality, and Interaction*, Lecture Notes in Computer Science, pp. 209–223. Springer.

Vranas, P. B. M. (2007). I ought, therefore I can. *Philosophical Studies 136*(2), 167–216.

Wang, Y. (forthcoming). A logic of goal-directed knowing how. *Synthese*. `https://doi.org/10.1007/s11229-016-1272-0`.

Weigend, T. (2014). Subjective elements of criminal liability. In M. D. Dubber and T. Hörnle (Eds.), *The Oxford Handbook of Criminal Law*, pp. 490–511. Oxford: Oxford University Press.

Wikipedia contributors (2017). Automated planning and scheduling. `https://en.wikipedia.org/w/index.php?title=Automated_planning_and_scheduling`.

Wilson, G. M. (1989). *The Intentionality of Human Action*. Stanford: Stanford University Press.

Wooldridge, M. (1997). Agent-based software engineering. *IEE Proceedings Software Engineering 144*(1), 26–37.

Wooldridge, M. and N. R. Jennings (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review 10*(2), 115–152.

von Wright, G. H. (1951). Deontic logic. *Mind 60*(237), 1–15.

Xu, M. (2015). Combinations of stit with ought and know. *Journal of Philosophical Logic 44*(6), 851–877.

Yu, E. (2001). Agent orientation as a modelling paradigm. *Wirtschaftsinformatik 43*(2), 123–132.

Samenvatting

Is het mogelijk dat een groep gezamenlijk verantwoordelijk is voor een bepaald resultaat terwijl geen enkel groepslid individueel verantwoordelijk is voor het bijdragen aan dit resultaat? Volgens een bekende paradox uit de socialekeuzetheorie is het verrassende antwoord ja. Stel je voor dat je solliciteert op een baan en na voorgeleid te zijn aan een sollicitatiecommissie het volgende te horen krijgt: de commissie heeft besloten je aan te nemen terwijl geen enkel commissielid afzonderlijk jou had willen aannemen. Dit soort gevallen worden *verantwoordelijkheidsgaten* genoemd.

Het blijkt dus dat verantwoordelijkheidsgaten voor kunnen komen, maar wat zijn nu precies de condities voor het ontstaan van dit soort lacunes? En, kunnen we iets zeggen over wanneer er wel een relatie is tussen collectieve en individuele plichten? In dit proefschrift zet ik op een systematische manier uiteen wat de relatie is tussen collectieve verantwoordelijkheid en de individuele verantwoordelijkheden van groepsleden. Door formele modellen uit de filosofie, economie en kunstmatige intelligentie te gebruiken en ontwikkelen ontstaat een genuanceerd beeld van deze relatie. Ik behandel hier dus geen specifieke casussen, maar zal een algemene theorie opzetten. Daarnaast wordt duidelijk dat er bepaalde condities zijn voor het ontstaan van de gevreesde verantwoordelijkheidsgaten.

Door deze relatie en deze condities bloot te leggen is het mogelijk om onze collectieve beslissingsprocessen te structuren zodat ze aan gewenste eigenschappen

voldoen. Dit draagt bij aan het verbeteren van collectieve beslissingsprocessen die essentieel zijn voor samenwerkingen, commissies, complexe systemen, en ook democratische instituties.

Voordat we dieper in deze kwestie duiken wil ik een spraakmakend voorbeeld behandelen: het discursieve dilemma. Stel je voor dat er een commissie is opgezet, bestaande uit drie leden, Marie, Mel en Mo, die moet beslissen of aan meneer Grensgeval een baan aangeboden moet worden op de universiteit. Laten we aannemen dat de universiteit hiervoor het beleid heeft dat een kandidaat excellent dient te zijn op het gebied van onderwijs, onderzoek én administratie. De commissie moet zijn oordeel bepalen door eerst over iedere afzonderlijke competentie te stemmen, vervolgens deze stemmen te aggregeren tot een groepsbesluit door middel van het meerderheidsprincipe, en tenslotte de beslissing over het aanbieden van een baan af te leiden aan de hand van het universiteitsbeleid. Als de leden van de commissie stemmen in lijn met de Figuur 1, dan is het resultaat dat ze gezamenlijk besluiten om de baan aan te bieden terwijl ze unaniem tegen dat besluit zijn. Stel je nu voor dat meneer Grensgeval helemaal geen goede kandidaat is. Is het in dit geval zinvol om te zeggen dat de commissie *collectief* verantwoordelijk is voor het aanbieden van de baan? Is een van de leden *individueel* verantwoordelijk voor het bijdragen aan deze onwenselijke uitkomst?

	Onderzoek z	Onderwijs w	Administratie a	Baan? $z \,\&\, w \,\&\, a$
Marie	Ja	Ja	Nee	Nee
Mel	Nee	Ja	Ja	Nee
Mo	Ja	Nee	Ja	Nee
Groep	Ja	Ja	Ja	→Ja / ↓Nee

Figuur 1: *Het discursieve dilemma.*

Dit voorbeeld (en andere soortgelijke voorbeelden) zijn recentelijk gebruikt om te argumenteren dat het zinvol is om te spreken over *collectieve* verantwoordelijkheid, naast individuele verantwoordelijkheid. Daarnaast toont dit voor-

beeld aan dat collectieve verantwoordelijkheid niet te reduceren is tot individuele verantwoordelijkheid. Met andere woorden, het laat zien dat collectieve verantwoordelijkheid nodig is, omdat individuele verantwoordelijkheid onvoldoende is om de verantwoordelijkheden in dit soort scenario's te ondervangen. Zonder collectieve verantwoordelijkheid zou er namelijk geen verantwoordelijke zijn.

Nu kun je denken dat de commissie uit het voorbeeld gewoon een foutieve collectieve beslissingsprocedure gebruikt. Echter, er zijn resultaten uit de sociale-keuzetheorie die aantonen dat geen enkele redelijke collectieve beslissingsprocedure dit soort gevallen uitsluit. Met andere woorden, iedere redelijke collectieve beslissingsprocedure biedt ruimte voor dit soort onwenselijk verantwoordelijkheidsgat. Er is dus geen enkele manier om verantwoordelijkheidsgaten uit te sluiten door enkel de collectieve beslissingsprocedure aan te passen.

Daarom is het belangrijk om te onderzoeken of er bepaalde condities zijn voor het ontstaan van de gevreesde verantwoordelijkheidsgaten. Door deze condities bloot te leggen is het mogelijk om onze collectieve beslissingsprocessen zo te ontwerpen dat ze aan gewenste eigenschappen voldoen. De bevindingen van dit proefschrift kunnen zo bijdragen aan collectieve beslissingsprocessen die essentieel zijn in, bijvoorbeeld, democratische instituties.

Om een antwoord te vinden op de hoofdvragen vormen twee ideeën de leidraad voor dit proefschrift. *Allereerst* is het van belang op te merken dat *handelingen* cruciaal zijn voor verantwoordelijkheid. Het is uiteraard vreemd om iemand verantwoordelijk te houden voor een bepaald resultaat wanneer diegene geen enkele handeling verricht heeft. De notie van handeling kan verschillend opgevat worden. Zo is er een duidelijk verschil tussen iets met opzet teweeg te brengen, het bewust teweegbrengen en het feitelijk teweegbrengen (waarbij de relevante kennis en intentie kan ontbreken). Deze *handelingsmodi* relateren aan drie verschillende niveaus van schuldigheid, bekend uit de rechtspraak: intentioneel handelen, bewust handelen, en causaal handelen.

Laat me deze verschillen verhelderen aan de hand van een voorbeeld. Stel je voor dat Amy, een jaloerse getrouwde vrouw, erachter komt dat haar man een affaire heeft met Vee. Met enkel het doel om Vee uit de buurt de verjagen, gaat ze op een nacht naar Vee's huis, belt aan om te na te gaan dat niemand thuis is, giet benzine over de voordeur en steekt die in brand. Vee komt om in de resulterende brand. Amy is gechoqueerd. Het kwam niet in haar op dat Vee mogelijk fysiek gevaar liep en ze had geen bewust plan om Vee om te brengen toen ze de deur in vlam zette. In dit geval is het nuttig om te onderscheiden dat Amy wel opzettelijk en bewust het huis in brand stak, maar niet opzettelijk of bewust Vee ombracht, terwijl ze dat laatste wel feitelijk teweegbracht.

We zouden dit verhaal kunnen aanpassen zodanig dat Amy wel wist dat Vee thuis was, terwijl Amy nog steeds alleen opzettelijk het huis in brand stak. Dit zou dan aantonen dat er een verschil is tussen bewust Vee ombrengen en opzettelijk Vee ombrengen.[1] Daarnaast kan de bewuste handelingsmodus ook verwijzen naar kennis van bepaalde omstandigheden, zoals in het bezit zijn van een kwaadaardige hond. De eigenaar van zo'n hond hoeft niet de intentie te hebben om iemand fysiek te schaden, maar zal wel aansprakelijk zijn voor het bewust fysiek schaden van omstanders wanneer deze worden aangevallen door zijn hond.

Ten tweede is het van belang in te zien dat de verantwoordelijkheden van een onafhankelijk individu verschillen van die van een groepslid. Zo kan het zijn dat ik als groepslid van een sportteam wel degelijk een plicht heb om aan alle trainingen deel te nemen, terwijl dat voor een onafhankelijk individu anders ligt. Het is dus belangrijk om deze twee *perspectieven* van elkaar te onderscheiden. Naast deze twee individuele perspectieven is er dan nog een *groeps*perspectief waarop de collectieve verantwoordelijkheden aanhaken.

[1] Desalniettemin is het gebruikelijk om te zeggen dat Amy roekeloos heeft gehandeld. Deze modus van roekeloos handelen laat ik in dit proefschrift buiten beschouwing.

Laat me deze twee individuele perspectieven verhelderen aan de hand van een klassiek publieke goederen probleem: het participatie dilemma. Stel je een gemeenschap voor die besloten heeft om een buurtwacht op te zetten om de criminaliteit terug te dringen. Laten we aannemen dat deze buurtwacht succesvol is dan en slechts dan als er een bepaald aantal leden participeert in het project. Uiteraard is het zo dat participeren kleine persoonlijke kosten teweegbrengt, maar dat een geslaagde buurtwacht een voordeel oplevert voor de gehele gemeenschap. De persoonlijke 'opbrengst' van een bepaalde buurtgenoot hangt dus af van het succes van de buurtwacht en van of ze zelf participeert. Wat moet zo'n lid van de gemeenschap doen?

Het intuïtieve idee is dat het verschil tussen het individuele en het gemeenschapsgerichte perspectief hem zit in dat een onafhankelijk individu zich afvraagt 'Wat moet *ik* doen?', terwijl een gemeenschapsgericht individu zich eerst afvraagt 'Wat moeten *wij* doen?' om vervolgens te bepalen hoe ze daar het beste aan kan bijdragen. Zonder op dit moment een volledige theorie te geven, is het goed om op te merken dat, vanuit het individuele perspectief, een individu zal participeren slechts dan als ze het voldoende waarschijnlijk acht dat zij doorslaggevend is voor het slagen van de buurtwacht. Vanuit het gemeenschapsgerichte perspectief, zal een individu participeren tenzij ze het voldoende waarschijnlijk acht dat haar bijdrage irrelevant is voor het slagen van het project. Kortom, een gemeenschapslid is meer positief geneigd om bij te dragen dan een onafhankelijk individu.

Deze twee centrale ideeën geven twee dimensies – handelingsmodi en perspectieven – en leveren negen combinaties op die samengevat worden in Figuur 2. Om invulling te geven aan deze twee ideeën zal ik modellen uit de filosofie, economie en kunstmatige intelligentie gebruiken en ontwikkelen. Met behulp van deze modellen kan haarfijn worden onderzocht wat de wisselwerking tussen verschillende cruciale concepten is, zoals de relatie tussen intentie, kennis en actie. Door deze interdisciplinariteit zijn de bevindingen relevant voor velerlei vakgebieden,

al is dit proefschrift vooral filosofisch, of fundamenteel, van aard en legt het zich toe op de filosofie van gezamenlijk handelen, participatie en collectieve verantwoordelijkheid. Deze twee centrale ideeën geven een basis om een systematische studie op te zetten naar de relatie tussen collectieve en individuele beslissingen.

Perspectief / *Modaliteit*	**Individueel**	**Collectief**	**Groepslid**
Causaal	Causale handeling	Collectieve causale actie	Causale bijdrage
Bewust	Bewuste handeling	Collectieve bewuste handeling	Bewust bijdragen
Opzettelijk	Intentionele handeling	Collectieve intentionele handeling	Intentioneel participeren

Figuur 2: *Handelingsmodi en perspectieven.*

Laat ik nog even uitweiden over de rol en de schoonheid van deze interdisciplinariteit. Collectieve verantwoordelijkheid en verantwoordelijkheidsgaten komen alleen voor in zogenaamde wederzijds afhankelijke keuzeproblemen. Dat wil zeggen, situaties waarin de resulterende uitkomst, en onze evaluatie daarvan, afhangen van de interactie tussen verschillende individuen. Binnen de microeconomie is het uiterst gebruikelijk om zogenaamde speltheoretische modellen te gebruiken om dit soort keuzeproblemen te modelleren. In hoofdstuk 1 leg ik een verband tussen deze modellen uit de speltheorie en modellen die gebruikt worden in de filosofie en kunstmatige intelligentie. Deze verbinding schept de mogelijkheid om onderwerpen en onderzoeken vanuit deze disciplines aan elkaar te koppelen. Zo zal ik bijvoorbeeld iets zeggen over de relatie tussen economisch onderzoek over sociale voorkeuren en samenwerking, en filosofisch onderzoek naar gezamenlijk handelen en participatie (§ 4.4.4).

In dit proefschrift gebruik ik de categorieën uit Figuur 2 om de relatie tussen collectieve en individuele verantwoordelijkheid en de condities voor verantwoor-

delijkheidsgaten te onderzoeken. In hoofdstuk 2 leg ik me voornamelijk toe op de causale modus van handelen en op het vervullen van verplichtingen. Zo zal blijken dat er geen algemene relatie tussen collectieve en individuele plichten bestaat, maar ontdek ik condities die verantwoordelijkheidsgaten uitsluiten.

In hoofdstuk 3 richt ik me op de modus van bewust handelen door me te richten op de subjectieve capaciteiten van agenten, ofwel hun praktische kennis, in tegenstelling tot de feitelijke capaciteiten. Het kan namelijk zo zijn dat een agent feitelijk in staat is om X teweeg te brengen zonder dat zij dit weet. Ik identificeer verschillende mogelijke redenen voor het ontbreken van de collectieve praktische kennis die noodzakelijk is voor het vervullen van een collectieve verplichting. Tenslotte toon ik aan dat wanneer een groep wel de betreffende collectieve praktische kennis bezit dan is een collectief falen altijd toe te kennen aan een groepslid dat bewust niet haar deel uitvoert. Daarenboven zal, in dit geval, dit groepslid bewust het risico lopen dat zijzelf de (co-)veroorzaker is van het collectieve falen.

In hoofdstuk 4 richt ik me op de modus van intentioneel handelen. Hierin verbind ik de filosofische theorieën over collectieve intentionaliteit en intentionele participatie aan economische theorieën over sociale voorkeuren en samenwerking door bestaande modellen uit te breiden met intenties. Vervolgens onderzoek ik twee gevallen. Allereerst bekijk ik het geval waarin een groep collectief intentioneel niet zijn collectieve verplichting vervult en toon ik aan dat dit impliceert dat alle leden intentioneel participeren in dit collectief falen. In dit geval is het zo dat collectieve verantwoordelijkheid impliceert dat *ieder* lid individueel verantwoordelijk is. Tenslotte, wanneer een groep niet collectief intentioneel zijn collectieve verplichting vervult dan impliceert dit dat er een lid is dat niet intentioneel participeert in de juiste groepshandeling.[2] In dit geval betoog ik dat er goede redenen zijn om te zeggen dat dit lid individueel verantwoordelijk is voor het afzien van

[2]Er is een verschil tussen intentioneel iemand beledigen en niet intentioneel voorkomen dat je iemand beledigt. Binnen de filosofie wordt dit verschil vaak verwoord als passief of actief iets teweegbrengen.

participeren in de juiste groepshandeling. De collectieve verantwoordelijkheid kan hier als het ware getraceerd worden naar dit individu. Wanneer zij goede redenen heeft om af te zien van participatie, kan het echter voorkomen dat ze verontschuldigd is voor dit gebrek.

In hoofdstuk 5 onderzoek ik de mogelijkheid dat verantwoordelijkheidsgaten ontstaan aan de hand van het verschil tussen samenwerking en competitie. Hiervoor gebruik ik een redenatie-gebaseerde aanpak. Dat wil zeggen, ik schets een theorie van verantwoordelijkheid op basis van het praktisch redeneren van de betreffende agent. In dat praktisch redeneren spelen causaliteit, verwachtingen en intenties een belangrijke rol – net zoals deze concepten een belangrijke rol spelen in de eerdergenoemde handelingsmodi. Vervolgens betoog ik dat competitieve gevallen vrij zijn van verantwoordelijkheidsgaten, terwijl coöperatieve gevallen mogelijk verantwoordelijkheidsgaten bevatten. Daarnaast geef ik condities waaraan voldaan moet zijn opdat verantwoordelijkheidsgaten kunnen voorkomen. Met andere woorden, ik geef condities die voldoende zijn om verantwoordelijkheidsgaten uit te sluiten.

Laat ik kort besluiten. Dit proefschrift gebruikt en ontwikkelt theorieën afkomstig uit de filosofie, economie en kunstmatige intelligentie om verantwoordelijkheidsgaten te onderzoeken. Ondanks dat de relatie tussen collectieve en individuele verantwoordelijkheden complex is, ontwikkel ik een systematische aanpak om de condities voor de ongewenste verantwoordelijkheidsgaten te identificeren. Op deze manier kunnen de bevindingen bijdragen aan collectieve beslissingsprocessen die essentieel zijn voor samenwerkingen, commissies, complexe systemen en democratische instituties.

Curriculum Vitae

Hein Duijf was born on 3 October 1989 in Maasbree, the Netherlands. He obtained his bachelor's and master's degree in Mathematics at Radboud University Nijmegen in 2011 and 2013, respectively, (the latter with the distinction *cum laude*) in which he specialized in Algebra and Logic by, for instance, doing a semester abroad in the Master Computational Intelligence at the Technical University of Vienna. He worked as a Junior Consultant at Aia Software between 2013 and 2014 and then started his PhD project at Utrecht University in the *Intelligent Systems* group in May 2014 on the research project titled 'Responsible Intelligent Systems' funded by the European Research Council. He finished his PhD project at the *Theoretical Philosophy* group at Utrecht University. In 2017 he was a visiting scholar at the London School of Economics for one term. Between 2015 and 2018 Hein was a member of the PhD Council of the Humanities faculty. He has taught courses on mathematics, logic, artificial intelligence, and philosophy of action and the social sciences. He has published in peer-reviewed conference proceedings (*Strategic Reasoning* and *Logic, Rationality and Interaction*) and academic journals (*Economics & Philosophy*, *Logique et Analyse*, *Philosophy of the Social Sciences*, and *Philosophical Studies*). His formal-philosophical studies embrace collective agency, intentionality, moral responsibility, practical reasoning, and interactive epistemology.

This page intentionally contains only this sentence.

Quaestiones Infinitae

PUBLICATIONS OF THE DEPARTMENT OF PHILOSOPHY AND RELIGIOUS STUDIES

VOLUME 21. D. VAN DALEN, *Torens en Fundamenten* (valedictory lecture), 1997.
VOLUME 22. J.A. BERGSTRA, W.J. FOKKINK, W.M.T. MENNEN, S.F.M. VAN VLIJMEN, *Spoorweglogica via EURIS*, 1997.
VOLUME 23. I.M. CROESE, *Simplicius on Continuous and Instantaneous Change* (dissertation), 1998.
VOLUME 24. M.J. HOLLENBERG, *Logic and Bisimulation* (dissertation), 1998.
VOLUME 25. C.H. LEIJENHORST, *Hobbes and the Aristotelians* (dissertation), 1998.
VOLUME 26. S.F.M. VAN VLIJMEN, *Algebraic Specification in Action* (dissertation), 1998.
VOLUME 27. M.F. VERWEIJ, *Preventive Medicine Between Obligation and Aspiration* (dissertation), 1998.
VOLUME 28. J.A. BERGSTRA, S.F.M. VAN VLIJMEN, *Theoretische Software-Engineering: kenmerken, faseringen en classificaties*, 1998.
VOLUME 29. A.G. WOUTERS, *Explanation Without A Cause* (dissertation), 1999.
VOLUME 30. M.M.S.K. SIE, *Responsibility, Blameworthy Action & Normative Disagreements* (dissertation), 1999.
VOLUME 31. M.S.P.R. VAN ATTEN, *Phenomenology of choice sequences* (dissertation), 1999.
VOLUME 32. V.N. STEBLETSOVA, *Algebras, Relations and Geometries (an equational perspective)* (dissertation), 2000.
VOLUME 33. A. VISSER, *Het Tekst Continuüm* (inaugural lecture), 2000.
VOLUME 34. H. ISHIGURO, *Can we speak about what cannot be said?* (public lecture), 2000.
VOLUME 35. W. HAAS, *Haltlosigkeit; Zwischen Sprache und Erfahrung* (dissertation), 2001.
VOLUME 36. R. POLI, *ALWIS: Ontology for knowledge engineers* (dissertation), 2001.
VOLUME 37. J. MANSFELD, *Platonische Briefschrijverij* (valedictory lecture), 2001.
VOLUME 37A. E.J. BOS, *The Correspondence between Descartes and Henricus Regius* (dissertation), 2002.
VOLUME 38. M. VAN OTEGEM, *A Bibliography of the Works of Descartes (1637-1704)* (dissertation), 2002.
VOLUME 39. B.E.K.J. GOOSSENS, *Edmund Husserl: Einleitung in die Philosophie: Vorlesungen 1922/23* (dissertation), 2003.
VOLUME 40. H.J.M. BROEKHUIJSE, *Het einde van de sociaaldemocratie* (dissertation), 2002.
VOLUME 41. P. RAVALLI, *Husserls Phänomenologie der Intersubjektivität in den Göttinger Jahren: Eine kritisch-historische Darstellung* (dissertation), 2003.
VOLUME 42. B. ALMOND, *The Midas Touch: Ethics, Science and our Human Future* (inaugural lecture), 2003.
VOLUME 43. M. DÜWELL, *Morele kennis: over de mogelijkheden van toegepaste ethiek* (inaugural lecture), 2003.
VOLUME 44. R.D.A. HENDRIKS, *Metamathematics in Coq* (dissertation), 2003.
VOLUME 45. TH. VERBEEK, E.J. BOS, J.M.M. VAN DE VEN, *The Correspondence of René Descartes: 1643*, 2003.
VOLUME 46. J.J.C. KUIPER, *Ideas and Explorations: Brouwer's Road to Intuitionism* (dissertation), 2004.

VOLUME 47. C.M. BEKKER, *Rechtvaardigheid, Onpartijdigheid, Gender en Sociale Diversiteit; Feministische filosofen over recht doen aan vrouwen en hun onderlinge verschillen* (dissertation), 2004.
VOLUME 48. A.A. LONG, *Epictetus on understanding and managing emotions* (public lecture), 2004.
VOLUME 49. J.J. JOOSTEN, *Interpretability formalized* (dissertation), 2004.
VOLUME 50. J.G. SIJMONS, *Phänomenologie und Idealismus: Analyse der Struktur und Methode der Philosophie Rudolf Steiners* (dissertation), 2005.
VOLUME 51. J.H. HOOGSTAD, *Time tracks* (dissertation), 2005.
VOLUME 52. M.A. VAN DEN HOVEN, *A Claim for Reasonable Morality* (dissertation), 2006.
VOLUME 53. C. VERMEULEN, *René Descartes, Specimina philosophiae: Introduction and Critical Edition* (dissertation), 2007.
VOLUME 54. R.G. MILLIKAN, *Learning Language without having a theory of mind* (inaugural lecture), 2007.
VOLUME 55. R.J.G. CLAASSEN, *The Market's Place in the Provision of Goods* (dissertation), 2008.
VOLUME 56. H.J.S. BRUGGINK, *Equivalence of Reductions in Higher-Order Rewriting* (dissertation), 2008.
VOLUME 57. A. KALIS, *Failures of agency* (dissertation), 2009.
VOLUME 58. S. GRAUMANN, *Assistierte Freiheit* (dissertation), 2009.
VOLUME 59. M. AALDERINK, *Philosophy, Scientific Knowledge, and Concept Formation in Geulincx and Descartes* (dissertation), 2010.
VOLUME 60. I.M. CONRADIE, *Seneca in his cultural and literary context: Selected moral letters on the body* (dissertation), 2010.
VOLUME 61. C. VAN SIJL, *Stoic Philosophy and the Exegesis of Myth* (dissertation), 2010.
VOLUME 62. J.M.I.M. LEO, *The Logical Structure of Relations* (dissertation), 2010.
VOLUME 63 M.S.A. VAN HOUTE, *Seneca's theology in its philosophical context* (dissertation), 2010.
VOLUME 64 F.A. BAKKER, *Three Studies in Epicurean Cosmology* (dissertation), 2010.
VOLUME 65 T. FOSSEN, *Political legitimacy and the pragmatic turn* (dissertation), 2011.
VOLUME 66 T. VISAK, *Killing happy animals. Explorations in utilitarian ethics.* (dissertation), 2011.
VOLUME 67 A. JOOSSE, *Why we need others: Platonic and Stoic models of friendship and self-understanding* (dissertation), 2011.
VOLUME 68 N. M. NIJSINGH, *Expanding newborn screening programmes and strengthening informed consent* (dissertation), 2012.
VOLUME 69 R. PEELS, *Believing Responsibly: Intellectual Obligations and Doxastic Excuses* (dissertation), 2012.
VOLUME 70 S. LUTZ, *Criteria of Empirical Significance* (dissertation), 2012
VOLUME 70A G.H. BOS, *Agential Self-consciousness, beyond conscious agency* (dissertation), 2013.
VOLUME 71 F.E. KALDEWAIJ, *The animal in morality: Justifying duties to animals in Kantian moral philosophy* (dissertation), 2013.
VOLUME 72 R.O. BUNING, *Henricus Reneri (1593-1639): Descartes' Quartermaster in Aristotelian Territory* (dissertation), 2013.
VOLUME 73 I.S. LÖWISCH, *Genealogy Composition in Response to Trauma: Gender and Memory in 1 Chronicles 1-9 and the Documentary Film 'My Life Part 2'* (dissertation), 2013.
VOLUME 74 A. EL KHAIRAT, *Contesting Boundaries: Satire in Contemporary Morocco* (dissertation), 2013.
VOLUME 75 A. KROM, *Not to be sneezed at. On the possibility of justifying infectious disease control by appealing to a mid-level harm principle* (dissertation), 2014.

VOLUME 76 Z. PALL, *Salafism in Lebanon: local and transnational resources* (dissertation), 2014.
VOLUME 77 D. WAHID, *Nurturing the Salafi Manhaj: A Study of Salafi Pesantrens in Contemporary Indonesia* (dissertation), 2014.
VOLUME 78 B.W.P VAN DEN BERG, *Speelruimte voor dialoog en verbeelding. Basisschoolleerlingen maken kennis met religieuze verhalen* (dissertation), 2014.
VOLUME 79 J.T. BERGHUIJS, *New Spirituality and Social Engagement* (dissertation), 2014.
VOLUME 80 A. WETTER, *Judging By Her. Reconfiguring Israel in Ruth, Esther and Judith* (dissertation), 2014.
VOLUME 81 J.M. MULDER, *Conceptual Realism. The Structure of Metaphysical Thought* (dissertation), 2014.
VOLUME 82 L.W.C. VAN LIT, *Eschatology and the World of Image in Suhrawardī and His Commentators* (dissertation), 2014.
VOLUME 83 P.L. LAMBERTZ, *Divisive matters. Aesthetic difference and authority in a Congolese spiritual movement 'from Japan'* (dissertation), 2015.
VOLUME 84 J.P. GOUDSMIT, *Intuitionistic Rules: Admissible Rules of Intermediate Logics* (dissertation), 2015.
VOLUME 85 E.T. FEIKEMA, *Still not at Ease: Corruption and Conflict of Interest in Hybrid Political Orders* (dissertation), 2015.
VOLUME 86 N. VAN MILTENBURG, *Freedom in Action* (dissertation), 2015.
VOLUME 86A P. COPPENS, *Seeing God in This world and the Otherworld: Crossing Boundaries in Sufi Commentaries on the Qurʾān* (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 ʿImād 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.