Thinking before Acting

Intentions, Logic, Rational Choice

Olivier Roy

Thinking before Acting

Intentions, Logic, Rational Choice

ILLC Dissertation Series DS-2008-03

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
phone: +31-20-525 6051
fax: +31-20-525 5206
e-mail: `illc@science.uva.nl`
homepage: `http://www.illc.uva.nl/`

Thinking before Acting
Intentions, Logic, Rational Choice

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D.C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Aula der Universiteit
op dinsdag 26 februari 2008, te 10.00 uur

door

Olivier Roy

geboren te Sept-Iles, Québec, Canada.

Promotiecommissie:

Promotor:
Prof.dr. J.F.A.K. van Benthem

Co-promotor:
Prof.dr. M.V.B.P.M. van Hees

Overige leden:
Prof. dr. M. Bratman
Prof. dr. R. Bradley
Prof. dr. K. R. Apt
Prof. dr. F. Veltman

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The investigations were supported by the Fond Québécois de la Recherche sur la Société et la Culture (bourse # 93233), and by the Conseil de la Recherche en Science Humaine du Canada (bourse # 752-2006-0345).

Cover design by Lin de Mol.
Original artwork on the back cover : '*Trá*', 2004 C-print, 20 x 30 cm.
Copyright © Lin de Mol, Courtesy Galery Lumen Travo

Printed and bound by Printpartners Ipskamp.

ISBN: 978–90–5776–174–4

À Christine, Chantal et Omer

Contents

Acknowledgments

First of all I want to thank my promotor Johan van Benthem. He provided me both the freedom and the support I needed to develop the ideas behind this thesis. His comments and suggestions were always extensive and inspiring, and came right when I needed them. I am especially grateful for his presence and availability during the writing phase of this dissertation. Each section of it has profited from his advices. I am also grateful to him for getting me involved in many of his projects—joint publications, organization, seminar and book edition. Through them I have learned a lot about academic life, and especially about how to do fruitful collaborative research. I also want to thank him for making possible two visits to Stanford University, which have been among the most intellectually stimulating periods I ever had.

With Martin van Hees I have developed in two papers—see [2007a; 2007b]—the ideas behind Chapter 2, 3 and 4. I am very grateful to him for his carefulness and patience; always reading, commenting and correcting in depth the numerous drafts that we exchanged in the last two years. He became my co-promotor relatively late in the process, but his guidance, always very thoughtful, proved extremely precious. He is also indirectly responsible for me handing in this thesis well on time: by offering me a post-doc position which will start in February 2008!

I want to thank Michael Bratman for introducing me to the themes and issues which now form the core of Chapter 6, and for taking the time to read, comment and discuss its various versions. Without Michael's sharp philosophical sense, openness and extensive knowledge of the literature, this chapter would have never grew to the present state from the eight pages note it was in spring 2006.

Richard Bradley's contribution to this thesis has been less direct, but nevertheless crucial. By commenting a talk I gave at the LSE in June 2005, he has been one of the first "external" reader of what would become my dissertation theme. His positive and encouraging remarks gave me the drive I needed to pursue further the research on formal approaches to intentions. During the same conference he also introduced me to Martin van Hees' work. A small pointer which shifted the course of my PhD.

The thesis profited from the precious comments and encouraging remarks of many others. Discussions with David Israel have inspired me a lot, and where always lively and enjoyable. I want also to thanks Michael Wooldridge, Giacomo Bonanno, Adam Brandenburger, Dov Samet, Branden Fitelson, Rohit Parikh, Alexandru Baltag, Krister Segerberg and Luc Bovins. Either via extended discussion or small but highly accurate remarks, they all have greatly influenced this work.

Many faculty members of the ILLC have also commented drafts or talks related to this thesis. I am very grateful to Robert van Rooij, Paul Dekker, Peter van Emde Boas, Ulle Endriss, Frank Veltman and Krzysztof Apt. Special thanks to the last two who, as members of my thesis committee, also when through the whole (uncorrected!) version of the manuscript. I would also like to acknowledge the ILLC faculty members who taught the various courses I attended during the first year: Maricarmen Martinez, Yde Venema, Benedikt Löwe and Reinhard Blutner. Most of the technical results in this thesis build in one way or another from what they taught me. The ILLC staff also deserve many thanks for their support: Marjan Veldhuisen, Ingrid van Loon, Tanja Kassenaar, Peter van Ormondt and Karin Gigengack.

During the whole PhD period I have been involved in Johan's research on preference logic, first with Sieuwert van Otterloo and later with Patrick Girard. I have grown a lot from these fruitful collaborations and, more importantly, found dear friends. Many other fellow PhD's and postdocs at the ILLC have provided precious comments and suggestions for this work. First and foremost my friend and colleague Eric Pacuit, with whom I had so many enjoyable discussions, and from whom I received so many valuable advices. Thanks also to Fenrong Liu, for all the help and comments in the last years, to Jonathan Zvesper, Cedric Dégremont, Sujata Ghosh, Yanjing Wang, Floris Roelofsen, Fernando Velazquez-Quesada and Paul Harrenstein.

The circle of friends and colleagues who also contributed to this work extends way beyond the ILLC, and all deserve my most grateful thanks. Tomasz Sadzik and Paul Egré, also dear friends and colleagues. Mikaël Cozic, especially for his quite heroic comment on my paper at a workshop in Paris... the day before handing in his own PhD thesis! Many thanks to Marc Pauly, Conrad Heilmann, Rosja Mastop, Barteld Kooi, Boudewijn de Bruin, Martin Bentzen, Denis Bonnay, Cedric Paternotte, Tomohiro Hoshi, Randall Harp, Facundo Alonso, Jennifer Morton, Sarah Paul and Darko Sarenac.

My friends in Amsterdam may not have directly contributed to the content of this thesis, but without them I would not have brought it to the final stages. Jelle and Nick, above all. Diederik, flatmate and friend. Merlijn and Loes, to San Francisco! Lauren's always refreshing presence. Those with whom I shared the excitements and anxieties of the first year in Amsterdam: Maxim, Gaëlle, Scott

(maximum respect!), Höskuldur, Chiaki, Stéphanie, Edgar, Ka-Wo and Stefan. Good friends all over the town: Paul Janssen, Marian, Jan-Willem, Leigh, Jill, Tania, Kim, Lin, Aram, Sasha, Eva, Franz, Tjasa, Bjorn, Melle, Curtis and Karen. Special thanks to Lin, for designing the cover design and giving me the permission to reproduce her work on it. Thanks to those belonging to the "(extended) third floor" over the years: René, Joost, Brian, Reut, Aline, Yoav, Clemens, Marco, Khalil, Sara, Joel, Chantal, Fabrice, Michael, Tikitu, Hartmut, Daisuke, Andreas, Nina, Jakub, Levan, Jacob, Olivia, Yurii, Amélie, Raul, Anouk, Alessandra, Lena, Gideon and David Chabot. Special thanks to Joel, for the amazingly thorough check of the typesetting for the very last version of the manuscript. Thanks also to those I (almost) only know in swimsuit: Jim, Alexander, Jasper, Eric, Karl-Michael, Ana-Lara, Ben and Wendy.

Back in Québec, I want to thank Daniel Vanderveken, who supervised my master thesis, encouraged me to come to Amsterdam and invited me to present my work there in a conference in May 2007. Many thanks to Michel, Simon, Candida, Carlos and Florian for the enjoyable discussions and useful comments. I and very grateful to Mathieu Marion, Renée Bilodeau, Denis Fisette, François Lepage, Pierre-Yves Bonin and Claude Thérien for their support over the years.

Finally, love to Omer, Chantal and Christine and Hanna.

Source of the chapters

The ideas developed in Chapter 2 stem from [van Hees and Roy, 2007b]. Those in chapter 3 and 4 from [van Hees and Roy, 2007a]. Chapter 5 builds on joint work with Johan van Benthem, Sieuwert van Otterloo and Patrick Girard, spread in [van Otterloo, 2005], [van Benthem et al., 2005], [van Benthem et al., Forthcoming] and some unpublished notes.

The financial support of Fond Québécois de la Recherche sur la Société et la Culture (bourse # 93233) and of the Conseil de la Recherche en Science Humaine du Canada (bourse # 752-2006-0345) is gratefully acknowledged.

The spelling and grammar of this thesis (except these acknowledgments!) have been corrected by Ian Priestnall from *Paragraph Services*.

Chapter 1

Introduction

Practical reasoning is generally seen as the “human capacity for resolving, through reflection, the question of what one is to do.” [Wallace, 2003b] It is the process through which agents balance *reasons* for actions and eventually make their decisions. Among the many theories of practical reasoning, very few have such a long history as that of *instrumental rationality* , which goes back at least to Aristotle[1], and is still one of the most prominent modes in theoretical economics, especially in *decision* and *game theory*[2]. Practical reasoning, from the instrumental point of view, is a matter of finding the best means to achieve one's ends. In modern vocabulary, it is a matter of determining the actions which have the best expected consequences in terms of one's *preferences*.

Contemporary approaches to rational decision making have shed new light on this very ancient notion. Work on the foundations of decision theory, for instance, has given precision to the conditions under which one can view the choices of an agent in uncertain situations as instrumentally rational. Similarly, recent results in what is called the *epistemic program* in game theory[3] have highlighted the importance of mutual expectations in understanding rational behaviour in strategic interaction.

Arguably, much of this progress has been achieved thanks to increasing interest in the decision maker's *information*. In most contemporary analyses, *knowledge* and *belief* occupy just as important a place as preferences, to the extent that some decision and game theorists now hold that any adequate description of a decision situation should include a model of the agent's information. The development of these models, in turn, owes much to contemporary *dynamic epistemic*

[1] See *Nichomachean Ethics* [III, 1112b, 10].

[2] von Neumann and Morgenstern [1944], Savage [1954] and Jeffrey [1965] are classics in these fields and have inspired much of the work in this thesis. More contemporary textbooks include Myerson [1991], Osborne and Rubinstein [1994], Binmore [2005] and Osborne [2004].

[3] For more references on the topic see [Aumann, 1999], de Bruin [2004], Brandenburger [2007] and Bonanno [2007].

logic[4]. This extensive analytical toolbox, at the intersection of philosophical logic and theoretical computer science, was designed precisely to analyse *reasoning* about information, higher-order information - viz., information about one's own or others' information - and information change.

In short, modern theories of instrumental rationality rest on two pillars: preferences and information. As such they offer quite a different picture from the one displayed in contemporary philosophy of action[5], where great emphasis is laid on the fact that *intentions* are as important as preferences and information in practical reasoning.

Historically, intentions have been seen as *outputs* of deliberation. As Broome [2002, p.1] puts it, "intending to act is as close to acting as reasoning alone can get us." From this point of view, practical reasoning is "reasoning that concludes in an intention." Many philosophers of action, in contrast, have emphasised that intentions are also active *inputs* to deliberation, *on a par* with preferences and information, and *irreducible* to them. They stress that human agents are not only rational but also *planning* agents, that is agents who deliberate and form intentions *in advance*, before the time for action comes. For such agents, *future-directed* intentions foster coordination, "filter" the set of admissible options into new deliberations and steer practical reasoning towards "relevant" means. In short, contemporary theories of intentions conceive practical reasoning as a matter of weighting options, *against a background of future-directed intentions.*

In this thesis I propose a theory that does justice to this idea but at the same time capitalizes on the resources of contemporary theories of instrumental rationality and dynamic epistemic logic in order to obtain a more all-encompassing picture of practical reasoning for planning agents. I show that such a broad approach genuinely enriches existing models of rational decision making, as well as the philosophical theory of intentions.

To develop this theory I draw on the contemporary paradigms of reasoning just mentioned: *decision* and *game theory* from theoretical economics, the *planning theory of intentions* from philosophy of action and *dynamic epistemic logic* from philosophy and computer science. In the following pages I briefly introduce them in an informal manner. What is important for now is to grasp the main ideas and intuitions. The formal details will come later, as needed in the subsequent chapters.

[4] Hintikka [1962] is the classic reference in epistemic logic. Fagin et al. [1995] presents a more up to date overview of the topic. For *dynamic* epistemic logic see Plaza [1989], Gerbrandy [1999], Baltag et al. [1998] and van Ditmarsch et al. [2007]. On the dynamic of information in games, see van Benthem [1999, 2001, 2005, 2006b] and van Benthem et al. [2005].

[5] Key references here are Davidson [1980], Harman [1976], Bratman [1987, 1999, 2006a] Velleman [2003] and Wallace [2006]. A good survey is available in the form of the reader on philosophy of action edited by Mele [1997].

1.1 Instrumental rationality, decision and game theory

Decision and game theory are two branches of theoretical economics which, as stated, conceive rational decision making as *instrumental rationality*. That is, they model practical reasoning as the process of discovering which actions are the best means to achieve one's ends.

1.1.1 Decision theory: rational decision making under uncertainty

Decision theory models *individual* decision makers who have to choose between a series of options, usually thought of as *actions*, *plans* or *strategies*. These decision makers are assumed to have some *ranking* over the set of options, which is usually represented as a comparison relation or a choice function[6]. This ranking is supposed to reflect the way the agents would choose if they were offered a choice between pairs or sets of options. For example, if action A and action B are options for the agent, then to say that action A is ranked above action B means that the agent would choose A if he had to decide between doing A and doing B.

When there is no uncertainty, it is quite clear when such choices are instrumentally rational. Provided that each action yields a certain *outcome*, about which the agent has some *preferences*, being instrumentally rational is just chosing the actions which yield a most preferred outcome. That is, an agent is instrumentally rational when he chooses action A over B whenever he prefers the outcome A yields to the outcome yielded by B.

Representation results in decision theory give conditions under which one can view the agent's choices *as if* they were those of an instrumentally rational decision maker. These results show how to *construct* or *extract* a preference relation over the set of outcomes from the ranking of actions, usually in terms of *payoffs* or *utilities*, in such a way that the agent chooses an action if and only if it yields an outcome that is most preferred. In more "behaviouristic" terms, a representation result shows how to reconstruct the agent's preferences over outcomes, which are essentially internal mental states, on an external or observable basis, namely his choices[7].

For decision situations without uncertainty it is clear how this construction should proceed. The most interesting situations, though, at least from a decision-theoretic point of view, are those *with* uncertainty. Two kinds of uncertainty are

[6] In this thesis I work only with relations. For more references on choice functions Sen [2002, chap.3].

[7] This behaviouristic stance is especially clear in Savage [1954, p.17] : "I think it of great importance that preference, and indifference, [...] be determined, at least in principle, by decision between acts and not by response to introspective questions."

considered in decision theory: *exogenous* and *endogenous*.

Situations of *exogenous* uncertainty are situations in which the results of the agent's actions depend on *random* or *non-deterministic* occurrences in the environment. Exogenous uncertainty is thus the result of *objective* non-determinacy, external to the decision maker. Buying a lottery ticket is a typical example. Someone cannot simply choose to buy "the" winning ticket. All one can do is pick one and wait for the drawing to determine whether it is a winner. In models of decision making under exogenous uncertainly, the options become "lotteries" or simply *probability distributions* over the set of outcomes. Buying a lottery ticket, for instance, would then be represented as an action which gives the agent a certain probability of winning.

Situations of *endogenous* uncertainty, on the other hand, are situations in which the outcomes of the agent's decisions depend on the *actual state of the world*, about which he has only partial information. The following is a classical example[8]. Imagine an agent who is making an omelet. Suppose he has already broken five eggs into the bowl and he is about to break the sixth and last one. Whether this will result in a good omelet depends on the state of this last egg, which can be rotten or not. The agent, however, does not know whether the egg is rotten. In slightly more technical jargon, he lacks information about the actual state of the egg, and *this* makes the outcome of his decision uncertain. Note the contrast with buying a lottery ticket. There is no chance or randomness involved here. The egg is rotten or it is not. What matters is that the agent does not know the state of the egg. Decision theoretic models, for example the one proposed by Anscombe and Aumann [1963], usually represent such situations by adding a set of possible *states* to the model. In the omelet example there would be two states, "the egg is rotten" and "the egg is not rotten". The uncertainty over these states is represented either *qualitatively*, by an epistemic accessibility relation, or *quantitatively*, by a probability distribution.

In uncertain situations it is less clear what the instrumentally rational choices are, because each decision has many possible consequences which the agent might value differently. Maximization of *expected* payoffs is the most widely accepted expression of instrumental rationality in such contexts[9]. The expected payoff of an action is the sum of the payoffs of all its possible outcomes, weighted by their respective (objective and/or subjective) probability. To be instrumentally rational, in the decision theoretic sense, is to choose the action, or one of the actions, which maximizes this sum. In other words, the agent is instrumentally rational when he chooses an action which gives the best prospect over outcomes.

Representation results for decision making under uncertainty provide condi-

[8] This example comes from Savage [1954, p.14].

[9] Rationality as maximization is not the only criterion for rational choice or instrumental rationality in decision theory. The maximin or minimax rules are other well known alternatives. The interested reader can consult any classical decision theory textbook, such as [Luce and Raiffa, 1957]. In what follows, however, I shall stick to rationality as maximization.

tions under which the agent's choices can be seen as the expressions of instrumental rationality, in this sense. They show that if the agent's choices of action satisfy certain stated conditions, then one can interpret the agent's decisions as a maximization of expected payoff. I shall not go into the details of these conditions here; they would lead us too far afield[10]. The point is that decision theory is a theory of instrumental rationality. It explains the conditions under which the choices of an individual decision maker facing uncertain options can be regarded as instrumentally rational.

1.1.2 Game theory: reasoning with mutual expectations

Game theory is also a theory of instrumental rationality, but now in contexts of *strategic interaction*, i.e., when the outcome of a decision situation is determined by the decision, not of *one* but of *many* rational decision makers. Game theory is usually divided into *cooperative* and *non-cooperative* branches. Here I am interested mostly in the non-cooperative side, although I briefly revisit the cooperative branch in Chapter 3. Nor do I address the "evolutionary" branch of game theory here, either.

The basic ingredients of a situation of strategic interaction, or simply of a *game*, are rather similar to those in decision theory. Each agent chooses among various *actions*, *plans* or *strategies*, the outcomes of which are uncertain. The crucial difference from decision-theoretic contexts is that this uncertainty now comes from the choices of other rational decision makers[11]. Indeed, outcomes in games are determined by a *combination* of the choices of all agents.

For reason that will soon become clear, most game-theoretic modelling bypasses the "representation" step[12]. The agent's preferences over the set of outcomes are just taken as givens, rather than being extracted from the choice behaviour. It is also assumed that the agents are instrumentally rational, i.e. that they choose in order to maximize their expected payoffs. The whole difficulty is now to specify what *are* the expected payoffs of an action, when its outcome depends on the actions of others.

The fact is that agents might not be able to anticipate the choices of others. This can happen, first, because they are uncertain about each other's preferences. These are situations of *incomplete information*[13]. In this thesis I examine only cases of *complete* information, that is, games in which the agents know each others' preferences. Even in such cases the agents might not know each others'

[10] See von Neumann and Morgenstern [1944], Savage [1954], Anscombe and Aumann [1963] and Jeffrey [1965] for details.

[11] Nothing precludes the introduction of exogenous uncertainty in games, too. See e.g. Myerson [1991, p.38-46].

[12] The work of La Mura and Shoham [1998] is a notable exception.

[13] Games with incomplete information have been introduced by Harsanyi [1967-68]. See also Myerson [1991, p.67-73] for a general overview.

choices, because the structure of the game prevents them from doing so[14] or because they are choosing *simultaneously.* These are cases of *imperfect* information.

In games with imperfect information each agent forms expectations about the others' decisions before making his own. That is, each agent bases his choices on what he *expects* the others to do. Here the contrast with decision theory is quite instructive. In decision situations under exogenous uncertainly the decision maker's expectations are *fixed* by the probability distribution induced by the objective random event. Similarly, in decision situations under endogenous uncertainty the decision maker's expectations are *fixed* by the (partial) information he has about the state of the world. These expectations are fixed in the sense that they do not depend on the agent's preferences and possible choices. The situation is fundamentally different in games. Each agent tries to anticipate what the others will do before making his decision. But each agent also knows that the other will do the same. This means that an agent's expectations about the others' actions take into account the fact that the others choose on the basis of what they think he will do. But theses choices are, in turn, made after taking into account the fact that he takes the choices of others into account. The bulk of modern game theory, which stems mostly from von Neumann and Morgenstern [1944] and Nash [1950], is precisely aimed at understanding what maximization of *expected* payoffs means with such a "circular" or potentially infinite regress of mutual expectations.

The solution of this problem is obviously less clear-cut than that for single-agent decision making. There is nothing in game theory like the strong decision theoretic consensus around "one" expression of instrumental rationality. Rather, expected payoff maximization in games is multifaceted, as witnessed by the great variety of *solution concepts*, the most well known of which are probably *iterated elimination of dominated strategies* and *Nash equilibrium.* Each different solution concept has corresponding, different expectations, which *epistemic characterizations* make explicit[15]. The strategies which are iteratively not dominated, for example, can be shown to be strategies which *rational* agents play when it is *common knowledge* that all agents are rational. That is, when all agents expect the others to be rational and to have reciprocal expectations. More generally, epistemic characterization results explain game-theoretical solution concepts in terms of the agents' mutual *information*, i.e. first-order knowledge and beliefs about each others' preferences and choices and higher-order knowledge and beliefs about the knowledge and beliefs of others.

Arguably, mutual expectations are to games what choices of uncertain actions are to decision theory. They provide the raw material in which the various conceptions of instrumental rationality are phrased. From that point of view, epistemic

[14] Consider, for example, the game "Battleship", in which neither player can see how the other has positioned his/her fleet on the game board.

[15] I return to epistemic characterizations of solution concepts, with more references, in Chapter 3.

characterizations of solution concepts can been seen as the natural game-theoretic counterparts of the decision theoretical representation results.

Decision and game theory are thus two theories of instrumental rationality. The first helps to understand conditions under which a single decision maker can be seen as instrumentally rational when he chooses among uncertain options. The second also provides conditions for seeing a decision maker as instrumentally rational, but this time as a function of how his expectations interlock with those of other rational agents. There is nothing, in both cases, which precludes one from seeing decision making as resulting in future-directed intentions. In the *process* of decision making, though, such intentions play no role. The "background of deliberation" in decision and game theory is constituted by preferences and information. There are no previously adopted intentions there, at least not of the kind which have been studied in philosophy of action.

1.2 Intentions and planning agency

In contemporary philosophy of action, the work of M. Bratman [1987, 1999, 2006a] is probably the most coherent and complete theory of intentions. Furthermore, even though it is not formal, Bratman's precise style of analysis has fostered dialogue with more formal approaches, especially in AI and computer science[16]. For these two reasons, most of the presentation here draws from his work.

For a long time, the dominant philosophical theories of intentions were *reductionist* and most attention was directed towards intentions *in action* . Intentions in action are the "mental components" [O'Shaughnessy, 1973] of intentional action, they are "what is left over [when] I subtract the fact that my arm goes up from the fact that I raise my arm" [Wittgenstein, 2001, § 621]. This distinguishes intentions in action from *future-directed* intentions, which are intentions to do something or to achieve some state of affairs *later.* Future-directed intentions essentially precede intentional actions, while intentions in action accompany them. Reductionist theories deny that intention, whether in action or future-directed, is a basic kind of mental state. Rather, they see intentions as compounds of more basic states, generally beliefs and desires[17]. To act with some intention, or to have the intention to do something later is, from that point of view, to have the appropriate belief-desire pair concerning that present or future action.

Bratman's [1987] proposal is *non*-reductionist and gives priority to *future-directed* intentions. It is built on the assumption that human agents are not only rational decision makers but also *planning* agents. Planning agents are agents who can "settle in advance on more or less complex plans concerning the future, and

[16] See for example Cohen and Levesque [1990] and Bratman et al. [1991].

[17] D. Davidson [1980] has been very influential in this tradition. See also Velleman [2003] and Setiya [2007].

then [let] these plans guide [their] later conduct." [*idem*, p.2] In short, planning agents are agents who can form future-directed intentions.

Most work on the *planning theory of intention*, including Bratman's, is *functionalist*. Future-directed intentions are described in terms of the regularities "which connect [them] with each others, with associated psychological processes and activities, and with characteristic "inputs" and "outputs." [*idem*, p.9]. Following a similar distinction proposed by Schwitzgebel [2006] for beliefs, these regularities can be divided into *backward* and *forward* looking.

The backward-looking ones concern the way intentions are "formed, revised and extinguished" [Shah and Velleman, forthcoming]. On this issue, I take the planning theory to be roughly in line with the traditional view of intentions: they are typically formed as the upshot of practical reasoning[18]. Future-directed intentions are states the agent gets into once he has settled on a particular course of action. Similarly, intentions are revised or given up on the basis of further deliberations. This makes them *relatively stable*, especially in comparison with desires. Preferences or desires change quite often and, more importantly, these changes need not be based on deliberations. An agent who made the decision to act in a certain way seems, on the other hand, to be committed to sticking to this decision by the reasons which led to it, unless counterbalancing reasons appear and trigger further deliberations. In other words, intentions are relatively resistant to *reconsiderations*.

The adjective "relative" is very important here. An agent should not constantly reconsider all his previous decisions, but neither should he stick to them at any cost. Rational intentions should be open to revision. But *how* open is a delicate—and crucial—question for the planning theory. I shall not go into detail on this issue. Intention revision occupies a relatively minor place in this thesis[19]. Most of the considerations are rather focused on the forward-looking characterization of intentions, which concerns their place in practical reasoning. Intentions influence this process by two forms of commitment that they carry: the *volitive* and the *reasoning-centered commitment.*

To say that intentions carry a volitive commitment is to say that they have some motivational force. Just like desires, they are metal states which push the agent to act, which trigger action. They are "conduct controlling" [Bratman, 1987, p.16]. An important part of the non-reductionist argument in the planning theory rests on this volitive commitment. Future-directed intentions carry a stronger motivational force than desires. An agent can have the desire to accomplish a certain action without ever enacting it. But if he genuinely intends to act in a certain way and does not revise this intention, he *will* normally do it in due

[18] I say "roughly" here because some authors argue that it is important to distinguish the result of a choice or a decision from the formation of a full-blown intention. See e.g. Bratman [1987, p.155-163].

[19] See the extensive discussion in [Bratman, 1987, chap.4-5-6], Harman [1986] and van der Hoek et al. [2007].

course.

It is a central claim in the planning theory that, because of their relative stability and their motivational force, intentions *anchor personal and inter-personal coordination.* Personal coordination is a matter of anticipating one's own decisions and actions, and of planning accordingly. An agent who, for example, has decided to pursue a PhD abroad for the next couple of years will probably take this future-directed intention into account while planning his next summer vacation. Maybe he will consider spending his vacation at home, something he would otherwise have not considered. But if intentions would normally not translate into action, or would most of the time be abandoned before their achievement, it would not be very useful for the agent to plan on the assumption that he will be abroad next year. The volitive commitment and the relative stability of intentions, from that point of view, provide a solid basis for temporally extended planning[20]. These considerations apply with even greater strength in *interaction.* Planning agents who are aware of each other's intentions are better capable of coordinating their actions, because they can rely on each other to do what they intend.

The volitive commitment of intentions already makes them valuable additions to the background of deliberations. In the words of Wallace [2006, p.105], they "resolve practical questions about my own agency," thus allowing planning agents to include them as "facts" into further planning. But previously adopted intentions are not confined to such a background appearance in practical reasoning. They also constrain and trigger deliberations, and by doing so they carry a reasoning-centered commitment. This second function stems from *norms* of consistency and coherence that apply to intentions.

Intentions are, first, required to be *means-end coherent.* This idea is very old, going back at least to Kant's hypothetical imperative[21]. Intentions about ends should be supplemented, at some point, with intentions regarding what the agent believes are necessary means to achieve these ends. An agent who intends to obtain his PhD in the next four years should at some point form more precise intentions on how he is going to make this happen. If he never forms such intentions, and does not even plan to "think about it", he is means-end incoherent.

Means-end coherence "poses problem for further deliberation" [Bratman, 1987, p.33]. It puts pressure on planning agents to deliberate about means to achieve what they intend. This is the first manifestation of the reasoning-centered commitment: intentions focus and even trigger new deliberations to find relevant means for their achievement.

[20] They also provide solid anchors for action in situations of temporary preference changes. In the words of McClennen [1990], they make planning agents more "resolute". By the same token, intentions also help to "break ties" between equally desirable options. I write more on this last point in Chapter 2.

[21] See Johnson [2004] or the original [Kant, 1785].

There is a second aspect of reasoning-centered commitment, which comes from the following three norms of consistency: *internal consistency*, *strong belief consistency* and *agglomerativity*. Agents have internally consistent intentions when they do not intend plain contradictions. Their intentions are strongly consistent with their beliefs[22] provided these intentions would be realizable in a world where the agent's beliefs are true. In other words, intentions are strongly belief consistent when they are capable of achievement, given what the agent believes. One can, finally, distinguish two understandings of agglomerativity. First, one can require the intentions of an agent to be closed under conjunction: if the agent intends to do A and intends to do B, then he should also intends to do A and B. Some authors find this requirement too strong, however. They argue instead that it should be possible to agglomerate intentions without violating other norms of consistency. That means, first, that if the agent intends to do A and intends to do B, then A and B should not be contradictory. Their conjunction would generate an internally inconsistent intention. Similarly, doing A and B should be consistent with the agent's beliefs.

Because of these norms of consistency, previously adopted intentions impose a "filter of admissibility" [Bratman, 1987, p.33] on the options that are to be considered in practical reasoning. The agent should be able to decide in favour of either of these options without generating, maybe after agglomeration, (belief) inconsistent intentions. Intentions simply rule out of deliberation such options as do not meet this requirement[23].

Before going any further, I should write a few words about the notion of "plan". The planning theory conceives of plans as *sets* of intentions with particular features[24]. In accordance with the means-end coherence requirement, plans are assumed to have a *hierarchical structure*. For whatever general intention a plan might contain, for example the intention to go to Paris, one should be able to find in the plan some subordinated means-intentions, for example going by train, depart at 9.15, and so on. The intentions at the bottom of this hierarchy, the most detailed ones, need not, however, settle every detail of the achievement of the plan. It is arguably even counterintuitive to suppose that they would, especially for agents with limited time and memory. Plans are typically *partial*, and planning agents fill them as needed. In the words of Savage, planning agents "cross the bridge when [they] come to it." [1954, p.16]

The functional characterization of intentions extends to plans. First, plans are formed, revised and abandoned as the upshot of practical reasoning. They

[22] The appellation "strong" comes from Bratman [1987]. Even though the name might suggest such, there is to my knowledge no "weak" version of this norm in the planning theory.

[23] It should be clear that I do not consider intention revision here. When discovering potentially inconsistent options, a planning agent might reconsider his intentions instead of simply ruling out these options.

[24] In this thesis, most of the time I simply use "intentions sets" rather than plans.

are also relatively resistant to reconsideration and commit the agent to their accomplishment. In virtue of being so they anchor personal and inter-personal coordination. Finally, they take part in further practical reasoning by triggering and focusing deliberations on means, and by imposing a filter of admissibility on the options to be considered.

1.3 Dynamic epistemic logic

I have already stressed many times the importance of *information* in the decision- and game-theoretic analyses of instrumental rationality. In each case rational choices are conditional on the agents' expectations, which are in turn very dependent on what they know and believe about the decision situation. Although I have put less emphasis on this aspect, information is also of great importance in the theory of intentions, especially for intentions in situations of interaction. When many agents are involved, what they know and believe about the intentions of others makes a crucial difference to what they intend and how these intentions influence further practical reasoning.

In that context, modern *dynamic epistemic logic* appears to be the natural environment for developing a theory of practical reasoning for rational planning agents. It is indeed a general theory of *reasoning* about information and information changes. It has already proved quite helpful in investigating the flow of information in games[25], and its tools and insights are readily applicable to questions of rational planning agency.

Modern dynamic epistemic logic is a branch of *modal* logic[26], a general framework devised to "talk", using *formal languages*, about some *structures* or *models.* As we shall see in the course of this thesis, decision and game-theoretic representations of decision situations are typical kinds of structures which modal languages can talk about.

Logical languages are connected to the structures that they are intended to describe via a *semantics*, which provides the conditions for the statements of the language to be *true* or *satisfied* in a structure. Semantics also allows one to isolate formulas that are *valid* in certain class of structures. These formulas are of special importance because they *correspond*, in modal terms, to general properties of the intended class of structures.

Alongside the semantic part, modal languages usually come with a deductive apparatus, called a *proof systems*. These allow explicit representation of the *reasoning* which can be done in a given formal language. These proof systems come in many forms, but in this thesis I use only *axiomatic* ones. These are

[25] See the references in the footnote on page 2.

[26] Blackburn et al. [2001] is a clear and up-to-date textbook on the topic. See also Blackburn et al. [2006].

constituted by sets of *axioms*, which are simply formulas in the language, and *inference rules*, with which one can syntactically derive *theorems*.

The semantic and deductive components are usually connected via *soundness* and *completeness* results. The first are intended to show that all axioms and derivable theorems are valid formulas in a given class of structure, and the second are intended to show that all the valid formulas in a given class of structure are either axioms or derivable from the axioms in the deductive system. If one can show that a given proof system is sound and complete with respect to a certain class of structure, one has shown that the deductive and the semantic component of the logical analysis exactly match. All valid principles about these structures are derivable, and all derivable statements are valid principles. As such, soundness and completeness results are extremely powerful analytical tools. With them, one can go back and forth between semantic and syntactic or deductive analysis. Any new valid formula that one finds has some corresponding deduction or reasoning in the proof system, and vice-versa. All formulas that one derives in the proof system are certain to be a valid principles describing general properties of the intended class of structures.

Epistemic logic uses these modal resources in order to study reasoning about information. Its formal language typically includes *knowledge* and *belief* operators, allowing one to form statements like "agent i knows that p" or "if agent i knows that p then agent j knows that q". Furthermore, by *combining* these operators epistemic languages can talk about *higher*-order information, that is knowledge and beliefs of agents about their own information, or the information of others.

The structures that epistemic logic talks about are usually called *epistemic models*, and they are very close to the qualitative models for endogenous uncertainty in decision and game theory. Roughly, they encode the agents' information using a given set of *states*, interconnected by some *relations*. These relations connect to a given state all the states that agents are not able to distinguish from the current one. In other words, they connect all the states that the agents consider possible. With such simple representation, when provided with some semantic definitions, modal languages are able to talk about surprisingly intricate epistemic conditions, involving arbitrarily high (finite) orders of knowledge and beliefs. What is more, they can establish a clear connection between conditions on the underlying relations and "properties" of the epistemic operators. Transitivity—which says that if state w is related to w' and w' to w'', then w is related to w''—can be shown, for instance, to correspond to "positive introspection" of the knowledge operator: if an agent knows that p then he knows that he knows that p. In other words, one can link the properties of an intended class of epistemic models with important valid principles about the agents' information in the epistemic language.

Epistemic logic also comes with well-known, sound and complete axiom systems for various classes of models and diverse epistemic operators. These allow one to neatly capture reasoning about information and higher-order information, and to connect it with valid principles. That it can do so in such a perspicuous manner is, arguably, the greatest asset of epistemic logic and what makes it so relevant to the analysis of practical reasoning.

Dynamic epistemic logic has been developed as an extension of epistemic logic to talk about information *changes*. More precisely, it is designed to talk about the consequences of "epistemic actions" on the information that agents have in a given situation. An epistemic action is simply an event which affects what the agents know, believe and consider possible. Dynamic epistemic logic has analyzed a whole variety of such actions, from simple public announcement of facts to hidden messages and encrypted communication.

These actions are usually represented by various *operations* which transform epistemic models, by removing states, by adding new ones or by changing the relations. Dynamic epistemic languages are extensions of epistemic languages, with operators corresponding to these operations. Again, if one is provided with a semantics, the properties of these operations can be shown to correspond to valid principles in the dynamic language[27]. In other words, general properties of information changes in epistemic models can be translated into valid principles about epistemic actions.

As for its epistemic base, dynamic epistemic logic comes with known, sound and complete proof systems in which one can study reasoning about information changes. These proofs systems are of special interest because most of them show how to analyze in a *compositional way* valid principles about information change in terms of principles about (static) information and higher-order information[28]. In other words, they show how to understand the effects of epistemic actions in terms of what the agents know and believe about each other in a given epistemic model.

In view of this, dynamic epistemic logic appears to be *the* natural environment to develop a theory of practical reasoning for rational planning agents. For one thing, the importance of information and higher-order information for rational decision making in itself justifies the use of epistemic languages and models. But, as I mentioned in Section 1.2, intentions bring in a key *dynamic* component, because they carry a reasoning-centered commitment through which agents transform their decision problems. As we shall see in Chapter 5, dynamic epistemic languages can readily be adapted to talk about this aspect of planning agency, and especially to study how intentions-based transformation of decision problems affects the agents' information in strategic games.

In this section I have focused on *epistemic* languages, that is, on modal lan-

[27] Studies on this correspondence are relatively recent in the field. See van Benthem and Liu [2004] and van Benthem [2007, Manuscript] for more details.

[28] A notable exception is the work of Gerbrandy et al. [2007].

guages which talk about information and information changes. But at this point it is worth mentioning that dynamic modal languages have also proved quite useful in the study of reasoning about *preferences* and preferences *changes*[29], which are also crucial ingredients in practical reasoning of rational agents. In Chapter 5 I make extensive use of these languages for preferences. This shows even more powefully how natural the choice of dynamic epistemic logic or, more generally, dynamic modal logic is as a framework for the development of a theory of intention-based practical reasoning.

1.4 Outline of the thesis

I start in Chapter 2 by showing that the introduction of future-directed intentions does indeed broaden the explanatory scope of *decision*-theoretic models, i.e. of models of *individual* decision making under uncertainty. As the planning theory claims, the volitive commitment of future-directed intentions allows one to go beyond traditional decision-theoretic reasoning by "breaking ties" between equally desirable options, and thus provides a straightforward anchor for personal coordination.

Chapter 2 also serves as launching pad for the rest of the thesis. I introduce there the formal framework and fix important methodological boundaries. I explain, for instance, why I bracket questions of *resource-boundedness*, crucial for "real-world" agents, and also why I do not consider intentions with *counterfactual consequences*. The reason is, in short, that by leaving these issues aside one can work in a much simplified decision theoretic environment while nevertheless highlighting important contributions of intentions to practical reasoning. Furthermore, these simple models generalize naturally to situations of strategic *interaction*, which occupy most of the investigation.

Coordination is the first aspect of intention-based strategic interaction that I consider (Chapter 3). It is a natural point for the theory of intentions to meet the theory of games. The first makes strong claims about coordination, and the topic has attracted much attention in the second. I mostly look at "Hi-Lo" games, which have become a standard benchmark for game-theoretic accounts of coordination. I show that intentions do indeed anchor coordination in these games, in a way that naturally generalizes their "tie-breaking" effect in single agent contexts. I also show that this intention-based account offers a plausible alternative to another proposal in the game-theoretical literature. I leave Hi-Lo games only by the end of the chapter, where I look at how intentions can anchor coordination in the general case. This allows us to revisit important claims in the planning theory concerning "shared agency", and in particular to circumscribe better the extent of this phenomenon.

[29] See e.g. van Benthem et al. [2005], van Benthem et al. [Forthcoming], de Jongh and Liu [2006], Liu [2008] and Girard [2008].

All of this concerns the volitive commitment of intentions, and how they "appear" in the background of practical reasoning. In Chapter 4 I turn to the more "active" role of future-directed intentions, namely the two facets of the reasoning-centered commitment: the filtering of options and the focus on means. I show that they can be studied by means of two simple operations which transform decision- and game-theoretic models. These operations become especially interesting in contexts of strategic interaction, where they acquire an important social character that has not yet been studied in the planning theory.

In Chapter 5 I use dynamic epistemic logic to bring the considerations of the previous chapters under a single umbrella. This provides a unifying theory of rational deliberation for a planning agent. Not only does it encompass both the volitive and the reasoning-centered commitment of intentions, it also allows one to study how these two functions interact. I show, for instance, that an important aspect of the volitive commitment used to account for coordination with intentions has an echo in the filtering of options that I define in Chapter 4. This observation triggers a natural generalization of the idea of filtering, which takes into account the information that agents have about their own intentions and the intentions of others. By the end of the chapter I explore two other issues at the intersection of planning agency and instrumental rationality, namely the condition under which intention-based transformations of decision problems foster coordination and become "enabled" by the elimination of dominated strategies.

The framework I develop in Chapter 5 is also, literally, a theory of practical *reasoning* of planning agents. *Axiomatic proof systems* for dynamic epistemic logic are well known. By adapting them to the framework of games with intentions, one gives a concrete deductive character to the theory. The findings about the various conditions for coordination, and about how they relate to transformations of decision problem, can be turned into reasoning or proofs in these axiomatic systems. In short, dynamic epistemic logic brings the present proposal closer to a fully-fledged theory of intention-based, rational deliberation.

In Chapter 6 I look back at this theory from a philosophical point of view, and investigate the question of how the norms of consistency and coherence which apply to intentions can be explained. In contemporary philosophy of action there are two main takes on this issue, called the "cognitivist" and "agency" approaches. Here I explore an alternative one, *hybrid pragmatism*, which stands half-way between cognitivism and the agency approach. It is based on the notion of "acceptance in deliberation", a cognitive state which has so far attracted little attention. I argue that hybrid pragmatism is a plausible alternative to the two main contemporary approaches, and that its use of acceptances provides a more complete picture of how future-directed intentions make their way into practical reasoning. Looking at hybrid pragmatism and acceptances in deliberation also brings this thesis to a natural conclusion. On the one hand, they supply a solid philosophical basis to the theory of rational planning agency. On the other, they open new, exiting research directions, both from the philosophical and the formal point of view.

All in all, this thesis is an attempt to capture, in a single framework, three important takes on practical reasoning. First, the considerations of instrumental rationality that have been extensively studied within decision and game theory. Secondly, the idea that intentions partake in most deliberations for planning agents, that they provide anchors for personal and inter-personal coordination, focus deliberations on relevant means, and filter options. Thirdly, the importance of mutual, higher-order and changing information in deliberation. In short, I propose here a framework for *rational planning agency* in which decisions, driven by the demands of instrumental rationality, are nevertheless made against a background of previously adopted intentions. This perspective, I believe, not only widens the scope of decision and game theory, but also unveils new issues for the philosophy of action.

Chapter 2

Intentions and individual decision making

This chapter has two aims. First and foremost I investigate intention-based individual decision making. That is, I look at how individual decision makers can take into account both considerations of instrumental rationality and future-directed intentions in their deliberations. Second, the chapter introduces the formal framework that I use throughout the thesis. The results on intention-based rational decision making are thus interspersed with many methodological remarks on the formal approach.

In Section 2.1 I introduce *extensive* representations of decision problems. These models allow a fine-grained representation of future-directed intentions, in which one can distinguish intentions to act from intentions to reach outcomes. In Section 2.2 I show that this framework permits a precise formulation of the norms of consistency and coherence which apply to intentions, and allows us to to study their interrelations. Moreover, one can account for the fact that planning agents can "break ties" between equally desirable options, and thus that the agents can use their intentions to coordinate their own decisions through time. In Sections 2.3 and 2.4 I set two important methodological boundaries on the investigation. I do not, first of all, consider intentions with "counterfactual" consequences, nor do I consider resource-bounded agents. As I show in Section 2.5, this allows one to work with very simple models, *strategic* decision problems with intentions, where one can really "zoom in" on phenomena at the intersection of the planning agency and instrumental rationality. These models, moreover, generalize easily to the multi-agent case, and thus provide a natural opening into the subsequent chapters.

2.1 Extensive decision problems

Decision theory is about rational decision making in the face of uncertainty. It offers a plethora of models, all more or less inspired by the pioneer works of Ramsey [1926], von Neumann and Morgenstern [1944], Savage [1954] and Jeffrey [1965]. Here I work with perfect information, extensive decision problems with exogenous uncertainty[1]. Just as in Figure 2.1, extensive decision problems can be seen as trees. Each path in the tree, called a history, is a sequence of "moves" or "actions". I put these terms in quotes because some of these actions are not performed by the agent, but are random occurrences in his environment. In more multi-agent terms, in these trees the agent is playing against Nature. He has no control over Nature's choices, which are the source of exogenous uncertainties[2]. In figures the white dots are choices or decision points while the black ones are the random occurrences or "Nature's moves". So, in Figure 2.1, the first node is a choice point while the one that follows action *Go up* is a coin toss, i.e. a chance node.

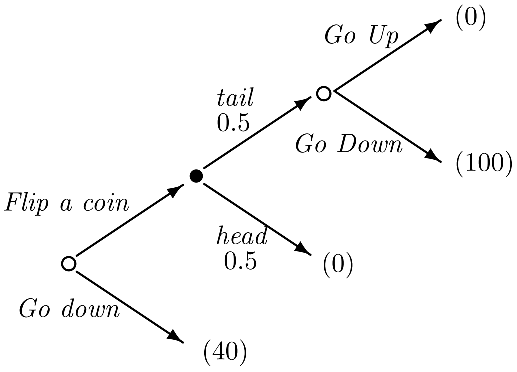

Figure 2.1: A simple decision problem

The sequential order of nodes in the tree, starting from the root and finishing at the leaves, represents the temporal succession of decisions. When the agent reaches a leaf, also called a terminal history or an outcome, he collects his payoff,

[1] Recall the distinction I made in the introduction, Section 1.1.1, between endogenous and exogenous uncertainty. In this chapter I do not consider endogenous uncertainty. It does, however, play an important role in Chapter 3, in the context of strategic games. See [Osborne and Rubinstein, 1994, Section III] for more on extensive decision problems with endogenous uncertainty.

[2] To push this idea a bit further, one could say that in decision trees without endogenous uncertainty, the agent "knows" Nature's (mixed, see the footnote on page 36) strategy. As I just mentioned, I do not look at cases of endogenous uncertainty, which would arguably include cases where the agent is uncertain about Nature's strategy, i.e. about the probability distributions associated with random events.

which I represent here as real numbers. The higher the payoff the better off the agent is. In Figure 2.1, if the agent reaches the terminal history *Go down*, that is if he chooses to go down at his first decision point, he collects a payoff of 40. He definitely would prefer to get 100, but this is not something he can enforce by himself. It depends on the result of the coin toss, an exogenously uncertain process over which he has no control.

Rational decision making in the face of such exogenous uncertainty is a matter of maximizing *expected* payoff. That is, a matter of maximizing the sum of the payoffs of all outcomes reachable by a sequence of choices, weighted by their respective probability. At this point it is useful to look at more formal definitions.

2.1.1. Definition. [Decision trees - Osborne and Rubinstein 1994] A *decision tree* T is a set of finite sequences of actions called *histories* such that:

- The empty sequence $\emptyset$, the *root* of the tree, is in T.
- T is closed under sub-sequences: if $(a_1, \dots, a_n, a_{n+1}) \in T$ then $(a_1, \dots, a_n) \in T$.

Given a history $h = (a_1, \dots, a_n)$, the history $(a_1, \dots, a_n, a)$, h followed by the action a, is denoted ha. A history h is *terminal in* T whenever it is the subsequence of no other history $h' \in T$. Z denotes the set of terminal histories in T.

2.1.2. Definition. [Chance and choice moves] The set of non-terminal history in a decision tree T is partitioned into two subsets, the *choice moves* and the *chance moves*. If h is a choice move, then the elements of $A(h) = \{a : ha \in T\}$ are called the *actions available at* h. If h is a chance move, then there is a probability distribution δ on $A(h)$, the elements of which are now called *alternatives at* h.

2.1.3. Definition. [Strategies and plans of action]

- A *strategy* s is a function that gives, for every choice move h, an action $a \in A(h)$. Equivalently, strategies are described as vectors of actions[3].
  - A node h' is *reachable* or *not excluded* by the strategy s from h if the agent can reach h' by choosing according to s from h. That is, h' is reachable by s from h if $h' = (h, s(h), s(s(h)), ...)$ for some (finite) application of s.
- A *plan of action* is a function that assigns to each choice node h that it does not itself exclude an action $a \in A(h)$. A *partial plan of action* is a partial function p' from the set of choice nodes to the set of actions which

[3] In Figure 2.1, for instance, there are four strategies: (Go down, Go Up), (Go down, Go Down), (Flip a coin, Go Up), (Flip a coin, Go Down).

coincides with a full plan of action[4]. A plan p *coincides* with a (perhaps partial) plan p' whenever $p' \subseteq p$.

A strategy tells the agent what to do at all choice points in the tree, even at those which are excluded by the strategy itself. In the example of Figure 2.1, (*Go down, Go Down*) is a strategy, even though by going down at the first node the agent would never reach the second or the third one. Plans of action, on the other hand, specify what to do only at choice nodes that are not excluded by the plan itself. Again in the figure above, (*Flip a coin, Go Down*) is the plan of first flipping the coin and then, if the coin lands tails, going down. Observe that many strategies can be compatible with a single plan of action. Going down at the first move in the example above excludes all other moves, and so (*Go down*) is a full plan of action compatible with both strategies (*Go down, Go Up*) and (*Go down, Go Down*). Partial plans are sets of decisions which could be extended to a full plan. In Figure 2.1 there are four partial plans: *Go down, Flip a coin, Go Up* and *Go Down.* The set of decisions $\{$*Go down, Go Down*$\}$ is not a partial plan in this sense. Going down at the first node excludes reaching the second choice point, and so there is no way to extend $\{$*Go down, Go Down*$\}$ to a full plan of action.

2.1.4. Definition. [Payoffs, reachable outcomes and expected payoffs]

- A *payoff* or *outcome function* π for a given decision tree T is a function that assigns a real number to each terminal history. In what follows it is assumed that every decision tree T comes with a finite set X of real numbers, where π takes its values.

- The set of *reachable outcomes* by a strategy s, or a full plan of action p, is the set of the $\pi(h)$ for all terminal histories reachable from the root by s (or p). The set of outcome reachable by a *partial* plan of action p is the union of the outcomes reachable by all the plans of actions p' that coincide with it[5].

[4] I should mention here an alternative way to define partial plans, used by van Benthem [2002]. Instead of defining them as partial functions, partial plans can be seen as total functions assigning *sets* of actions to each choice node. Partial plans, as I define them, boil down to functions that assign, at every history h, either a singleton $\{ha\}$—corresponding to cases where the partial function is defined—or the whole set $A(h)$—corresponding to cases where the partial function is not defined. This approach is truly more general, allowing "intermediate" cases where an agent has decided *not* to accomplish certain actions at a node, but has not yet made up his mind on which action, precisely, he will take. That is, cases were $p(h)$ is not a singleton but yet a strict subset of $A(h)$. The considerations in Section 2.2 would surely profit from this more general approach. I stick to partial plans as partial function, partly to keep things simple and partly to keep the presentation uniform with the notions of full plans and strategies.

[5] For convenience, I slightly abuse the notation and denote these sets by $\pi(s)$ or $\pi(p)$.

- The *expected value* or the *expected payoff* of a strategy s at the history h, denoted $EV(h, s)$, is defined inductively.

$$EV(h,s) = \begin{cases} \pi(h) & \text{If } h \text{ is a terminal history.} \\ \Sigma_{a \in A(h)} \delta(a) EV(ha, s) & \text{If } h \text{ is a chance node.} \\ EV(hs(h), s) & \text{If } h \text{ is a choice node.} \end{cases}$$

One can readily calculate that, in the example of Figure 2.1, the expected payoffs of the strategies (Go down, Go Up), (Go down, Go Down), (Flip a coin, Go Up), (Flip a coin, Go Down) are respectively 40, 40, 0 and 50. The expected value of a plan of action is computed similarly. Observe that, given Definition 2.1.4, a plan has exactly the same expected value at the root as all the strategies that coincide with it. The expected value of (Go down), for example, is the same as (Go down, Go Up), (Go down, Go Down).

I can now define more precisely what it means to be instrumentally rational at a node. It is simply to choose according to a strategy which expected value at that node is at least as high as that of any other strategy.

2.1.5. Definition. [Rationality] A strategy s is *rational at a node h* if and only if, for all strategies s': $EV(h, s) \geq EV(h, s')$. A strategy s is *rational for the whole decision problem T* if it is rational at the root of T.

The rational strategy of the decision problem in Figure 2.1 is thus (Flip a coin, Go Down).

Before introducing future directed intentions in these extensive representations, I should mention that, in this chapter, I am mostly concerned with plans of action and very little with strategies[6]. The latter are very important in multi-agent decision problems. The information they carry about "off-path" behaviour is crucial in Kreps and Wilson's [1982] *sequential equilibrium*, for instance. But in single agent scenarios, what the individual would do were he to deviate from his own plan of action has usually little or no consequence, at least as long as we are talking about ideal agents[7].

2.2 Actions- and outcomes-intentions in extensive decision problems

In this thesis I am mostly interested in cases where agents make their decisions against a background of previously adopted intentions. I thus do not consider

[6] With the exception of Section 2.3.

[7] I come back to the notion of ideal agent in Section 2.4. The reader might have already noticed that, in fact, I define rationality in a way that allows strategies to be rational even though they prescribe irrational moves at choice nodes that they exclude. In other words, the definition of what it is for a strategy to be rational in a decision problem already ignores the "off-path" prescriptions.

explicitly the process of intention formation[8], but rather focusing on cases where the agent comes to a decision problem with intentions that he has adopted beforehand.

An agent can come to a decision problem with both intentions to *accomplish certain actions* and intentions to *reach certain outcomes*. Let me call the first *actions*-intentions and the second *outcomes*-intentions. Extensive decision problems allow for a straightforward representation of both kinds of intentions.

2.2.1. Definition. [Intention sets] Given a decision tree T and a finite set of outcomes X, an *intention structure* $\mathcal{I}$ is a tuple $\langle I_O, I_A \rangle$ where I_O is a finite collection of sets of outcomes and I_A is a finite collection of (maybe partial) plans of action.

The elements in I_O are sets of outcomes. These are the outcome-intentions of the agent, intentions to achieve *some* of the outcome they contain. If, for example, outcomes x and y are outcomes where the agent gets more than 100 euros, then if $\{x, y\}$ is in I_O I will say that the agent intends to obtain more than 100 euros.

With this set-theoretical structure one can talk about various Boolean combinations of intentions. For instance, if A is a set of outcomes in which the agent is attending a concert, and B is a set of outcomes in which the agent is in Paris, then when both A and B are in I_O the agent intends to watch a concert and intends to be in Paris. If the intersection $A \cap B$ is in I_O then the agent intends to watch a concert *and* to be in Paris. Similarly, if $A \cup B$ is in I_O then he intends to watch a concert *or* to be in Paris. If, finally, the relative complement of A in X, noted $X - A$, is in I_O then the agent intends not to watch a concert. Note that this is different from the agent not intending to watch a concert, which is represented as $A \notin I_O$.

The set I_A is the set of action-intentions of the agent. If, for instance, the agent comes to the decision problem of Figure 2.1 with *Go Down* in his intention set, this means that he has already formed the intention to choose down if he reaches the second choice point.

So far there is no constraint on either I_A or I_O. It could be, for example, that I_O contains two mutually exclusive sets A and B. This would boil down to the agents having two contradictory intentions, since the achievement of one precludes the achievement of the other. Nor is there a relation between actions- and outcomes-intentions. An agent could intend to achieve outcome x without having any action-intention which would make him reach this outcome.

These are cases where the agent seems to violate one of the normative requirements on rational intentions. Recall that, for many philosophers of action,

[8] Introducing explicitly intention-formation points in decision trees also brings in important conceptual difficulties, mostly related to cases like G. Kavka's [1983] "Toxin Puzzle". I do not consider them in this thesis. M. van Hees and I have considered these questions in [2007b], using a model similar to the one of Verbeek [2002].

rational intentions should be internally consistent, consistent with what the agent believes, agglomerative and means-end coherent. With the exception of belief consistency, one can straightforwardly translate these norms in terms of intention structures[9].

2.2.2. Definition. [Internal consistency of outcomes-intentions] A set of outcomes-intentions I_O is *internally consistent* whenever $\emptyset \notin I_O$ and $I_O \neq \emptyset$.

If I_O contains the empty set the agent intends to achieve something utterly impossible, for example a plain contradiction. If, on the other hand, $I_O = \emptyset$ then the agent does not intend anything. Observe that this is different from intending no outcome in particular, which boils down to stating that I_O contains only the full set of outcomes, i.e. $I_O = \{X\}$.

2.2.3. Definition. [Agglomerativity of outcomes-intentions] The outcomes-intentions of the agent are *agglomerative* if $A \cap B \in I_O$ whenever $A \in I_O$ and $B \in I_O$.

This notion is essentially what I called agglomerativity *as closure* in the introduction (Section 1.2). Recall that I distinguished it from a another variant, agglomerativity *against potential irrationality*, according to which it should be possible to agglomerate the intentions of an agent without generating inconsistencies. In the present framework this alternative understanding of agglomerativity is a consequence of Definitions 2.2.2 and 2.2.3. Under the assumption that X and I_O are finite, internally consistent and agglomerative sets of outcomes-intentions always have a "smallest" element, which is never empty. In technical terms, $\bigcap_{X \in I_O} X \in I_O$ and $\bigcap_{X \in I_O} X \neq \emptyset$. I use a special notation for this intersection, the most precise outcomes-intention of the agent: $\downarrow\! I_O$.

For many of the technical details, it is convenient to assume that I_O is a *filter*, which means that it is not only internally consistent and agglomerative, but also closed under supersets: if A is in I_O and $A \subseteq B$ then B is also in I_O. Recalling the Boolean point of view on outcome-intentions, this closure under supersets is a closure under implication. If the agent intends to go to a Bob Dylan concert, then he intends to go to a concert. This is indeed a strong requirement, but it turns out to simplify the formal analysis[10].

Interestingly, closure under supersets has a completely different meaning with respect to the set of actions-intentions I_A. It rather expresses agglomerativity, both as closure and against potential irrationality.

[9] Recall that I work with perfect information decision trees. At each node the agent has certain and truthful information about his situation. His intentions, as long as they are achievable, are thus by default belief consistent. I work with imperfect information, for strategic games, in Chapter 3 and 5.

[10] Very few of the results, however, rest crucially on this assumption, as the reader will be in position to judge as I go along.

2.2.4. Definition. [Agglomerativity of actions-intentions] The actions-intentions of the agent are *agglomerative* if $p, p' \in M_A$ implies that $p \cup p' \in I_A$.

To see this, recall that all elements of I_A are plans of action, either full or partial. But then to require $p \cup p'$ to be an element of I_A, whenever p and p' are, is to require this union also to be a plan of action. This means that p and p' should not each exclude the achievement of the other. Observe that, given agglomerativity, actions-intentions also contain a "most precise" element, $\bigcup_{p \in I_A} p$, which can, however, still be partial. I use $\uparrow I_A$ to denote this.

Internal consistency could similarly be imposed on I_A as a separate requirement. But internal consistency can be imposed on actions-intentions in a less direct way. There is so far no connection between I_O and I_A. If, however, we view the latter as the set of *means* the agent intends to take in order to achieve his *ends*, i.e. his outcome-intentions, then there should be such a connection, namely a form of means-end coherence.

2.2.5. Definition. [Means-end coherence] The intention structure $\mathcal{I}$ is *means-end coherent* if there is a $p \in I_A$ such that $\pi(p) \subseteq \downarrow I_O$.

An intention structure is thus means-end coherent when there is a plan of action among the agent's actions-intentions which can, *for sure* yield an intended outcome. In other words, the agent's intention structure is means-end coherent whenever he can enforce the achievement of his outcomes-intentions by enacting his actions-intentions.

This condition not only constrains I_A in terms of I_O, it also precludes the agent from intending outcomes that he cannot secure. Consider a variation of the decision problem presented in Figure 2.1, in which π(*Flip a coin*, *tail*, *Go down*) $= 200$ instead of 100. In this case the agent has a clear favourite outcome. He cannot, however, intend to realize it and be means-end coherent at the same time. That is, all intention structures in which $\downarrow I_O = \{200\}$ are means-end incoherent. There is no (partial) plan of action for the agent in which he would be able to secure this outcome. To do so the agent would have to have the power to reach (*flip a coin, tail*) by his own will. But he can get there only if the coin lands tails up, and he has no control over that.

In connection with the other requirements, means-end coherence has interesting consequences. First, it is notable that internal consistency of actions-intentions follows from internal consistency of outcomes-intentions in the presence of means-end coherence. In other words, I_A is never empty for means-end coherent intention structures with internally consistent outcome-intentions. Furthermore, if I_A is agglomerative then the most precise action-intention $\uparrow I_A$ itself enforces an intended outcome.

2.2.6. Fact. For any means-end coherent intention structure $\mathcal{I}$, if I_A is agglomerative then, $\pi(\uparrow I_A) \subseteq \downarrow I_O$.

Proof. Take any such $\mathcal{I}$. Observe that for any (perhaps partial) plans of action p and p', $p \subseteq p'$ implies that $\pi(p') \subseteq \pi(p)$. Now, because I_A is agglomerative we know that for all $p \in I_A$, $p \subseteq \uparrow I_A$. But then, by means-end coherence, we know that there is a $p \in I_A$ such that $\pi(p) \subseteq \downarrow I_O$, and thus that $\pi(\uparrow I_A) \subseteq \downarrow I_O$. ■

As a direct corollary, we obtain that means-end coherence and agglomerativity of the actions-intentions together enforce that all actions-intentions are "consistent" with the most precise outcome-intention $\downarrow I_O$.

2.2.7. Corollary. *For any means-end coherent intention structure $\mathcal{I}$, if I_A is agglomerative then $\pi(p) \cap \downarrow I_O \neq \emptyset$, for all $p \in I_A$.*

Observe, however, that even if all the above conditions are met, $\uparrow I_A$ can still be a partial plan. This is quite in line with an important idea in the philosophy of action, namely that plans are typically incomplete. Means-end consistency "only" requires one to secure some intended outcomes, and this can leave many choice nodes undecided. Once the outcomes-intentions are within reach, actions-intentions can remain silent.

Means-end coherence nevertheless establishes a strong connection between the intentions of the agent and the structure of the decision tree. It constrains the intention structure to fit the agent's powers, keeping him within the limits of what he can enforce[11]. But one might also require the agent's intentions somehow to match another important component of decision problems, his *preferences*. That is, one might want to make sure that the agent's intentions will not lead him into choices that would otherwise be irrational. This can be ensured by the following.

2.2.8. Definition. [Payoff-compatible actions-intentions] The actions-intentions I_A are *payoff-compatible* whenever for any full plans of action p and p' such that $EV(\emptyset, p) > EV(\emptyset, p')$: if $\uparrow I_A \not\subseteq p$, then $\uparrow I_A \not\subseteq p'$.

2.2.9. Fact. For any decision tree T and means-end coherent intention structure $\mathcal{I}$ where I_O is agglomerative and internally consistent and where I_A is agglomerative and payoff-compatible, there is a rational plan that coincides with $\uparrow I_A$.

Proof. Simple unpacking of the definitions. ■

With this in hand we can already see that intentions extend the analysis of standard rational decision making under uncertainty. In cases where there is more than one maximizer of expected value, traditional decision theoretic agents have no criterion to decide. To put the matter somewhat dramatically, they are like Buridan's ass who, the story goes, starved to death because he could not decide between two equally desirable stacks of hay. As Bratman [1987, p.11] stresses,

[11] This condition, the "own action condition" [Bratman, 1999, p.148], will become very important in multi-agent scenarios. I discuss it in more detail at the beginning of Chapter 3.

planning agents *can* get out of cases like this, and Fact 2.2.9 is in line with this idea. When there is more than one rational plan in a decision problem solution, I_A can at most contain *one* of them, thus "focusing" traditional rationality. Moreover, when $\downarrow I_A$ is not a full plan of action, payoff-compatible intentions can still get the agent out of Buridan cases, by providing anchors for personal coordination. When there is more than one maximizer of expected value, some of which are not compatible with the agent's most precise action-intention, even if it is partial, a combination of traditional decision making and intention-based reasoning will help the agent to make a decision.

Intentions thus *supplement* traditional decision making under uncertainty, in a way that nevertheless does not fundamentally change the structure of the model. Actions- and outcomes-intentions are, after all, collections of entities that are "normal denizens" of extensive representations. This is an important asset of the present framework. By keeping the underlying decision-theoretic model intact, one can use the tools and insights from classical decision theory to gain a better picture of rational decision making with intentions. In other words, the current framework allows one to study intention-based rational deliberation in a way that does justice both to traditional views of instrumental rationality and the theory of intentions.

This essentially motivates the two methodological assumptions that I present in the following sections. Considering intentions with counterfactual consequences and resource-bounded agents introduces complications that lead us outside the core intersection of planning agency and instrumental rationality. These are interesting topics, but their investigation can surely profit from first looking at the "simple" case of rational deliberations against a background of future-directed intentions.

2.3 Intentions with counterfactual consequences

Consider the following example, in which the *counterfactual* consequences of intentions influence the payoffs[12].

> Both Terror Bomber and Strategic Bomber have the goal of promoting the war effort against Enemy. Each intends to pursue this goal by weakening the Enemy, and each intends to do that by dropping bombs. Terror Bomber's plan is to bomb the school in Enemy's territory, thereby killing children and terrorizing Enemy's population. Strategic Bomber's plan is [...] to bomb Enemy's munitions plant. [He] also knows, however, that next to the munitions plant is a school, and that when he bombs the plant he will also destroy the school, killing the children inside. [Bratman, 1987, p.139]

[12] In the literature such intentions are also called intentions with "double effects".

Let me assume that Strategic Bomber does not have the intention to kill the children, and try to draw the decision tree. If we make the standard assumption—as I have thus far—that a plan of action describes the available moves, then the tree is very simple (see Figure 2.2). There are only two possible plans—and also only two strategies—namely *bomb* and *not bomb*. But the consequences of *bomb* may be different if this history is chosen with the intention to kill the children than if it is chosen without this intention. For instance, Terror Bomber may be prosecuted for war crimes if it was indeed his intention to kill the children, whereas such prosecution may be less likely for Strategic Bomber. In this scenario, the payoffs thus not only depend on which terminal history is reached but also on the intention with which it is reached.

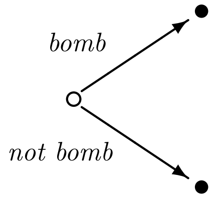

Figure 2.2: The Bombing Problem

In the model of the previous section, one cannot distinguish reaching *bomb* with the intention to kill the children from reaching the same history with a different intention. In both cases the agent has the same most precise actions-intention $\upharpoonright I_A = \{bomb\}$, and the value of $\pi(bomb)$ is the same, despite the fact that these intentions and payoffs should be different.

Bratman argues, in essence, that Strategic Bomber does not have the intention to kill the children because, in contrast to Terror Bomber, he *would* not adopt a new plan of action *if* the children were moved somewhere else, far from the munitions plant. That is, Bratman suggests that a *counterfactual extension* of the decision problem would reveal the intentions. One such extension is depicted in Figure 2.3. The first node is now a chance node which determines whether the school is deserted or not. If not, the agent faces the original decision problem, otherwise the counterfactual scenario arises. Considering this extended tree, one can assume that the plans of action of Terror and Strategic Bomber will differ. Terror Bomber's $\upharpoonright I_A$ will be the plan "Only bomb when children are at school" whereas for Strategic Bomber it will be "Always bomb". Following Bratman's suggestion, we can use the counterfactual information carried by these plans to assign the payoff to the terminal histories.

2.3.1. Definition. [Refined payoff functions] A *refined payoff function* ρ is a function that assigns a real-valued payoff to *pairs* (h, s) where s is a strategy and h a terminal history.

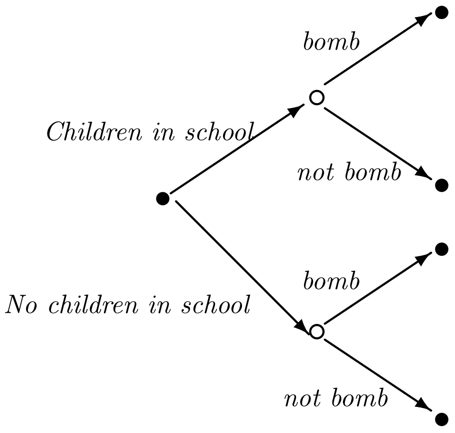

Figure 2.3: A counter-factual extension of the bombing problem

Dealing with refined payoff functions brings strategies back into full relevance. What the agent would choose were he to deviate from his own path, that is in counterfactual situations, suddenly has a direct impact on his payoffs. For that reason I will assume, for the moment, that actions-intentions can also contain strategies, with the corresponding modifications of Definitions 2.2.1 and 2.2.4.

2.3.2. Definition. [Intentions with counterfactual consequences] An actions-intention or a strategy s has *counterfactual consequences* when, $\rho(h, s) \neq \rho(h, s')$ for a terminal history h and another strategy s'.

In the Bombers example intentions have counterfactual consequences. The value of ρ is different if *Children in school, bomb* is reached with the strategy *Always bomb* than if it is reached with the strategy *Only bomb when children are at school.* Dropping bombs with the intention to kill the children has different consequences than dropping bombs without this intention.

The counterfactual consequences of acting with a certain intention can be taken into account when the agent chooses his strategy, with the following refinement of the notion of expected value.

2.3.3. Definition. [Refined expected value] For all pairs (h, s):

$$EV'(h,s) = \begin{cases} \rho(h,s) & \text{If } h \text{ is a terminal history.} \\ \Sigma_{a\in A(h)}\delta(a)EV'(ha,s) & \text{If } h \text{ is a chance node.} \\ EV'(hs(h),s) & \text{If } h \text{ is a choice node.} \end{cases}$$

Refined payoff functions are indeed generalizations of the standard payoff functions π. The latter are just refined functions for which the value at a terminal history is not dependent on the strategy with which it is reached. That is, any standard payoff function π can be emulated by a refined payoff function ρ for which $\rho(h, s) = \rho(h, s')$ for all terminal histories h and strategies s and s'.

However useful, refined payoff functions are not part of the standard decision-theoretic machinery. Two crucial questions remain open about them. First, I am not aware of any representation result that would ground a theory of instrumental rationality in uncertain situations, using these refined payoff functions. In other words, one still has to find the conditions under which choices of actions can be represented by such real-valued refined payoff functions on outcomes[13]. Second, one must understand better what kind of transformation would lead to the "correct" or "appropriate" counterfactual extension of a given decision tree. The use of refined payoff functions should come hand-in-hand with a systematic theory of decision problem transformations.

I do no purse these matters here. They are very interesting, but they bring in cases where intention-based practical reasoning forces one to reconsider the very building blocks of the theory of instrumental rationality. As we have already seen, even in "classical" decision-theoretic models intentions give rise to phenomena that are worth looking at, and to which I give priority.

2.4 Ideal and resources-bounded agents

So far I have considered examples where it is easy to find a rational plan. But this is not always so easy. If there are numerous choice nodes, interlaced with chance nodes, representing the decision tree or calculating its solution might be very tedious. Most decision-theoretic analyses abstract from such difficulties by making two assumptions about the agent: ideal *intelligence* and ideal *rationality.*

The first assumption concerns the agent's representational and computational capacities. An agent "is intelligent if he knows everything that we [the modeler] know about the [problem] and he can make any inferences about the situation that we can make." [Myerson, 1991, p.4] In other words, if a decision problem is representable at all and its solution computable, in any sensible sense of these terms, then the agent is assumed to be capable of representing it and computing its solution. The time and energy costs of these computations are simply ignored.

The rationality assumption splits into two components. First, the choices of the agent over strategies are assumed to satisfy certain "coherence" requirements, such as transitivity, totality and what Savage [1954] calls the "sure-thing principle".[14] These, together with a few others, are sufficient conditions to represent the agent's strategy choices as a maximization of expected value[15]. Decision-theoretic

[13] I briefly introduced representation results in the Introduction (Section 1.1.1).

[14] Transitivity states that if x is ranked above y, and y is ranked above z, then x is ranked above z. Totality states that, for any x and y, either x is ranked above y, or y above x. Finally, the sure-thing principle stipulates that if A and B produce the same consequences whenever some event E occurs, then the agent's choice between A and B should only depend on the consequences of these two actions in the case E does not occur. See Joyce [2004] for details.

[15] See the references in the footnote on page 29. In Definition 2.1.1, I directly introduced preferences in these terms.

agents are also assumed to be constant and flawless maximizers, meaning that at every choice point they choose according to a rational strategy, and that they do not make mistakes, i.e. make irrational decisions.

Ideal decision theoretic agents are thus perfectly rational agents who can represent any decision problem, however big, and find a rational plan without effort. These are indeed extremely strong idealizations, and most of them are explicitly made as simplifying hypotheses[16].

Decision models for non-ideal or resources-bounded agents have been studied, for instance, by Simon [1982], Rubinstein [1998] or Gigerenzer and Selten [2002]. It is also an important claim in philosophy of action that intentions are useful for agents with limited capacity, because they filter the set of options and focus deliberation on relevant means. By doing so, it is claimed, they reduce the number of options to be considered, leaving the other decisions for later. In short, intentions simplify deliberation, so that the agents can "cross the bridge when they come to it".[17]

But for ideal agents there is no need to wait until they come to the bridge. They are capable of computing *in advance*, for any decision problem, a maximally detailed plan. Furthermore, it is assumed that their preferences satisfy the above coherence requirements, and so that they will not want to change their plan along the way through the decision tree[18]. They are perfectly capable of pushing to its extreme the "look before you leap" approach:

> Making an extreme idealization, [...] a person has only one decision to make in his whole life. He must, namely, decide how to life, and he might in principle do once and for all. [Savage, 1954, p.83]

This idealization bears consequences for the representation of intentions, too. Ideal agents have little to do with partial plans. To put it in the words of von Neumann and Morgenstern [1944, p.79], the only assumption needed for agents to be able to choose (and intend) full plans of action "is the intellectual one to be prepared with a rule for behavior for all eventuality." This, they say, "is an innocent assumption within the confines of a mathematical analysis," that is under ideal decision-theoretic conditions.

Whether or not one agrees with the "innocent" character of this assumption, the point remains. For ideal agents, there is nothing that stands in the way of choosing beforehand a full plan of action. In other words, for ideal agents one can safely assume that $\uparrow I_A$ is a full plan.

[16] See for example the remarks of Savage [1954, p.30] about the computation costs.

[17] To formally assess the importance of future-directed intentions for resource-bounded agents one would also need to introduce considerations of computational complexity, which I do not do in this thesis.

[18] The argument for this last point rests crucially on the "representation" results that I mentioned earlier.

This has important consequences for the present analysis. First of all, it makes intentions even more powerful tools to get planning agents out of Buridan cases. If an agent's most precise actions-intention is a full plan, which happens to be payoff compatible, then this intention breaks any possible tie.

But this idealization also has methodological advantages. As I now show, it allows one to move to from extensive decision problems to the simpler strategic representations, without losing track of the important effects of previously adopted intentions in planning agency.

2.5 From extensive to strategic representations

Led by the observation that ideal agents can decide in advance on a full plan of action, von Neumann and Morgenstern [1944, p.79-84] proposed what they called the final simplification of decision problems, namely the *strategic form* or *normal* representation. Here is R. Myerson's [1991, p.50] account of von Neumann and Morgenstern's idea, phrased in game-theoretic terms.

> If the players in the game are intelligent, then each player should be able to [...] determine his rational plan of action before the game begins. Thus [...] the actual play of the game is just a mechanistic process of implementing these strategies and determining the outcome [...]. That is, we can assume that all players make all substantive decisions [...] at the beginning of the game, [...] [which] is exactly described by the normal representations.

In short, strategic representations of extensive decision problems "ignore all questions of timing." [*idem*, p.47]. The whole apparatus of decision trees and histories is replaced by a simple set of options, the elements of which represent plans of action. The preferences are simplified accordingly. Instead of being computed from the payoffs on terminal histories, they are directly aligned to expected value.

2.5.1. Definition. [Strategic version of extensive decision problems] Given a decision tree T and a payoff function π, its *strategic version* $\mathbb{G}_T$ is a tuple $\langle S, X', \pi, \succeq\rangle$ such that :

- S is the set of *plans of action* in T.
- X', the set of *outcomes*, is defined as $\{\pi(p) : p \in S\}$.
- $\succeq$ is a *preference relation* on X such that:

$$\pi(p) \succeq \pi(p') \text{ iff } EV(\emptyset, p) \geq EV(\emptyset, p')$$

In the same spirit, one can transfer intention structures for extensive decision problems to intention structures for strategic versions.

2.5.2. Definition. [Strategic version of intention structures] Given an intention structure $\mathcal{I}$ for a decision tree T, its strategic version $\iota = \langle \iota_A, \iota_X \rangle$ is defined as $\iota_A = \uparrow I_A$ and $\iota_X = \{Y \subseteq X : \pi(\uparrow I_A) \subseteq Y\}$.

The new actions-intention ι_A is defined by taking the most precise actions-intention in the extensive decision problem. Recall that if we work with ideal agents, we can assume that $\uparrow I_A$ is a full plan, which ensures that there is a p in the set S of plans in the strategic version which corresponds to the new ι_A. The new outcome-intentions set ι_X is, in turn, defined by taking the filter generated by the set of outcomes reachable by the most precise actions-intention $\uparrow I_A$. From Fact 2.2.9, we know that this set is a subset of the most precise outcomes-intention in the original extensive decision problem. By taking the filter generated by $\uparrow I_A$, it is thus certain that $\downarrow I_X$ will be an element of ι_X. In fact, this ensures that all elements of I_X are in ι_X.[19] What is more, defining the new intention structure in this way naturally preserves means-end coherence[20]. This transfer from extensive to strategic forms also retains the tie-breaking effect of intentions. If the actions-intentions of the agent are payoff-compatible and $\uparrow I_A$ is a full plan of action then ι_A is one of the maximal elements in the strategic preference ordering $\succeq$. That is, for all $p \in S$ we know that $\iota_A \succeq p$.

We thus have a simpler way to represent decision problems, strategic versions, where intentions can also break ties between equally desirable outcomes. This simplification can be pushed one step further, by abstracting from the extensive representations altogether.

2.5.3. Definition. [Decision problems in strategic forms] A *decision problem in strategic form* $\mathbb{G}$ is a tuple $\langle S, X, \pi, \succeq \rangle$ such that :

- S is a finite set of *plans* or *strategies.*
- X is a finite set of *outcomes.*
- $\pi : S \rightarrow X$ is an *outcome function* that assigns to every action an outcome $x \in X$.
- $\succeq_i$ is a reflexive, transitive and total[21] preference relation on X. Its strict sub-relation $\succ$ is defined as $x \succ y$ iff $x \succeq y$ but $y \not\succeq x$.

These are the strategic decision problems that I use in the coming chapters. Clearly, strategic versions of extensive decision problems are decision problems in strategic form. The preference relation induced by the expected value is indeed

[19] There might, of course, be more outcomes-intentions in ι_X than in I_X, if $\pi(\uparrow I_A) \subset \downarrow I_X$.

[20] Observe that it also preserves internal consistency and agglomerativity - the latter trivially, because I take the filter generated by $\pi(\uparrow I_A)$.

[21] I defined totality and transitivity in the Footnote on page 29. Reflexivity simply means that $x \succeq x$ for all outcome $x \in X$.

reflexive, transitive and total. The converse is no longer true, however. The difficulty lies in the preferences. One can transfer $\succeq$ from outcomes to plans, in the obvious way. But is not the case that any such preferences on plans, even if transitive, reflexive and total, can be represented by a numerical payoff function π on outcomes in an extensive game such that $p \succeq p'$ if and only if the expected value of p is greater than or equal to the expected value of p'.

This looser connection with extensive decision problems is compensated by a gain in generality. Strategic decision problems naturally extend to multi-agent interactive situations, on which I concentrate in the following chapters. Furthermore, by moving to decision problems in strategic form one can do away with the actions-/outcomes-intentions distinction, thus simplifying the formal apparatus even further.

2.5.4. Definition. [Intentions in strategic decision problems] An intention set ι for a strategic decision problem $\mathbb{G}$ is a set of subsets of X. The intention set ι is:

- *internally consistent* if $\iota \neq \emptyset$ and $\emptyset \notin \iota$.
- *agglomerative* if $A \in \iota$ and $B \in \iota$ implies that $A \cap B \in \iota$.
- a *consistent filter* if it is internally consistent, agglomerative and closed under supersets.

Observe that by the finiteness of X, we automatically get that if ι is a consistent filter then $\downarrow \iota$, defined as for $\downarrow I_O$, is not empty and an element of ι. Actions-intentions are not completely lost in this new definition. The agent can be seen as intending the plans which lead to outcomes in his most precise intention[22].

Because of the looser connection with extensive decision problems, payoff-compatibility can no longer be phrased in terms of expected value. To keep intentions within the boundaries of classical rationality I have to reformulate this criterion. Here I use a formulation which, I think, is fairly intuitive. The agent will be said to have payoff-compatible intentions if, when all the elements of a set of outcome B are strictly better than all the elements of another set of outcome A, if he still does not intend B, then he does not intend $A \cup B$ either. In plain English, if the agent prefers Holland to France as a holiday destination, but does not intend to holiday in Holland, then he does not intend to holiday in France or in Holland.

[22] Supposing that a given strategic decision problem can be translated back into an extensive one, this definition of actions-intentions would however violate agglomerativity in most cases. There will be in general more than one plan of action which will lead to intended outcomes. By putting them all in the actions-intention set, their union will clearly not be a plan of action. One should look in more detail at the question of "reconstructing" intentions for an extensive decision problem from intentions in a strategic one. Since I work with strategic representation from now on, I leave this issue aside.

2.5.5. Definition. [Payoff-compatible intentions] The intention set ι is *payoff-compatible* whenever for for any $A, B \subseteq X$, if $A \cup B \in \iota$ and for all $x \in A$ and $y \in B$, $x \succ y$ then $A \in \iota$.

This new version of payoff-compatibility characterizes precisely the intention sets $\downarrow\iota$ the elements of which are among the most preferred outcomes.

2.5.6. Proposition. *For any strategic decision problem $\mathbb{G}$ and intention set ι which is a consistent filter, the following are equivalent.*

1. *ι is payoff-compatible.*
2. *For all $x \in\downarrow\iota$ and $y \in X$, $x \succeq y$.*

Proof. The proof is obvious from (2) to (1). From (1) to (2), let $C(\succeq) = \{x \in X : x \succeq y \text{ for all } y \in X\}$. I show that $C(\succeq) \in \iota$. This will be enough because ι is a consistent filter. Take $A = C(\succeq)$ and $B = X - A$. Observe that $A \cup B = X$ which means, that $X \in \iota$. By definition of $C(\succeq)$, we know that $x \succ y$ for all $x \in C(\succeq)$ and $y \notin C(\succeq)$, that is, for all $y \in B$. We can thus conclude that $C(\succeq) \in \iota$ from payoff-compatibility. ■

This means that, just as in extensive representations, intentions in strategic decision problems can break ties between equally desirable options, thus genuinely adding to the standard decision-theoretic apparatus.

2.6 Conclusion

This last result exemplifies well the kind of analysis I carry out in this thesis. It highlights the effects of previously adopted intentions in practical reasoning of rational agents, without moving away from the underlying classical notion of instrumental rationality. In particular, it shows that one can study the effects of intentions even in the extremely simplified environment of strategic decision problems. Even if we assume ideal agency and leave aside intentions with counterfactual consequences, it is insightful to introduce intentions in decision-theoretic models.

Of course, I could have carried the analysis further using extensive decision problems, especially by introducing endogenous uncertainties or imperfect information. But in the coming chapter I rather move to interaction situations. We shall see that, on the one hand, the intentions bring with them insights into the study of coordination, and thus contribute to the general understanding of rational interaction. Interactive situations and game theory, on the other hand, also shed light on the theory of intentions by unveiling important issues about the reasoning-centered commitment.

Chapter 3

Intentions and coordination in strategic games

This chapter describes how intentions can foster coordination. More precisely, I investigate how agents in strategic interactions can successfully coordinate their actions by taking into account what they intend and what they know about the others' intentions, choices and preferences.

The bulk of the chapter (Sections 3.2 to 3.5) focuses on coordination in a very specific type of strategic interaction, namely *Hi-Lo games.* Such games provide a simple setting within which to "test" the ability of intentions to foster coordination. What is more, these games have become a benchmark in the study of coordination in game theory. By showing that agents can coordinate in Hi-Lo games on the basis of their intentions, I will thus be able to situate the planning theory better in relation to other game-theoretical accounts of coordination.

Before looking at Hi-Lo games, however, in Section 3.1 I tackle a fundamental issue concerning the very content of intentions in interactive situations. The difficulty is that, in such contexts, the agents' powers are limited. What results from an agent's decision depends in general on what the others decide. If we allow the agents to form intentions about any outcome, they will more often than not have intentions that they cannot achieve on their own.

Reflections on whether agents can form such intentions lead to issues concerning the volitive commitment carried by intentions and the information that the agents involved in games have about each other. To capture these ideas I use, in Section 3.4, 3.5 and 3.7, *epistemic models* for games. They provide a natural environment within which one can unfold the intention-based account of coordination in Hi-Lo games, in much the same fashion as game-theoretical epistemic characterizations of solution concepts[1]. This sheds new light on the *Stackelberg heuristic*, another explanation of coordination proposed by Colman and Bacharach [1997].

[1] I briefly introduced epistemic characterizations of solution concepts in the Introduction (Section 1.1.2).

Hi-Lo games are left behind in the last section of this chapter, where I use ideas from Bratman [1999, chap.5] to provide an intention-based account of coordination that does not rest on the specific structure of Hi-Lo games. This allows for a more general perspective on sufficient conditions for coordination in strategic contexts, permitting coordination to be compared with "shared cooperative activity."

3.1 Intentions in strategic interactions

In single-agent contexts without uncertainty the choices of the decision maker suffice to determine a unique outcome. In other words, the agent is able to realize any outcome he wants or intends. In decision problems with uncertainty the situation is not fundamentally different. The agent's choices do not determine a unique outcome, but rather a probability distribution on the set of outcomes. This probability distribution, however, reflects the agent's uncertainty about facts that are *independent* of what he believes, prefers or intends. The agent's capacity to realize what he wants or intends is thus bounded only by his own uncertainty and by randomness in his environment.

This crucially distinguishes single-agent decision problems from situations of strategic interaction, or *games*, where the choices of *all* agents determine the outcome[2]. What results from the decision of an individual in games depends greatly on something he cannot control: the choices of others. This can be captured by the following generalization of the single-agent strategic decision problems that I introduced at the end of the previous chapter[3].

3.1.1. Definition. [Strategic games] A *strategic game* $\mathbb{G}$ is a tuple $\langle I, S_i, X, \pi, \succeq_i \rangle$ such that :

- I is a finite set of agents.
- S_i is a finite set of *actions* or *strategies* for i. A *strategy profile* $\sigma \in \Pi_{i \in I} S_i$ is a vector of strategies, one for each agent in I. The strategy s_i which i plays in the profile σ is noted $\sigma(i)$.
- X is a finite set of *outcomes*.
- $\pi : \Pi_{i \in I} S_i \rightarrow X$ is an *outcome function* that assigns to every strategy profile $\sigma \in \Pi_{i \in I} S_i$ an outcome $x \in X$. For convenience I use $\pi(s_i)$ to denote the set of outcomes that can result from the choice of s_i. Formally: $\pi(s_i) = \{x : x = \pi(s_i, \sigma_{j \neq i})$ for some $\sigma_{j \neq i} \in \Pi_{j \neq i} S_j\}$.

[2] As I mentioned in the Introduction, this is so for games without exogenous uncertainty. See the remarks in the footnote on page 5.

[3] In this thesis I leave *mixed* or *probabilistic* strategies aside. A pure strategy is just an element of a set S_i for one agent i. A mixed strategy is a probability distribution on S_i.

- $\succeq_i$ is a reflexive, transitive and total preference relation on X.

The definition of the outcome function π captures the idea that outcomes are determined by the choices of *all* agents. It does not take single strategies or plans as argument but rather strategy *profiles* that is, combinations of choices[4]. It is this crucial difference which makes it necessary to reconsider what it means, in games, to have intentions to achieve outcomes.

It is a very common intuition, which also recurs frequently in philosophy of action[5], that agents can only intend what they have the power to achieve. If one allows agents to have arbitrary outcomes-intentions in games, one quickly runs into examples that clash with this idea.

	Cinema	Restaurant
Cinema	together	alone
Restaurant	alone	together

Table 3.1: A coordination game.

Consider the *coordination game* of Table 3.1. There are two agents, the row and the column agent, which I call 1 and 2. They have agreed to meet but they have forgotten where. Each agent can either go to the cinema or the restaurant. It doesn't matter to either of them where they meet, as long as they succeed in coordinating their actions, that is as long as they end up together at the cinema or together at the restaurant.

Suppose now that 1 intends to achieve (Cinema-Cinema). That is, he intends that he and his friend choose to go to the cinema, even though he cannot settle the matter himself. Following the intuition that agents can only form intentions that they can realize, one would say that 1 wishes or hopes that 2 will choose the cinema, but not that he has a genuine intention involving the choice of his friend. In other words, if we assume that agents can only intend outcomes that they can achieve by themselves, such intentions would have to be ruled out.

By following this line, though—restricting the set of intentions that an agent can form in strategic games—we would turn away from an interesting aspect of interactive intention-based practical reasoning. In a series of papers, Bratman [1999, chap.5 to 8] argued that intentions of the form "*I* intend that *we* do A" are the building blocks of "shared cooperative agency." Intentions about specific outcomes have precisely this form. In the example above, for 1 to intend (Cinema-Cinema) is for him to intend something like "*we*—1 and 2—meet at the movie

[4] It should be clear that the single-agent strategic decision problems from the previous chapter are particular cases of strategic games, where I contains only one agent.

[5] See the discussion in Bratman [1999, pp. 148-150], Baier [1970] and Velleman [1997]. The idea is even present in Aristotle, who wrote "we choose only what we believe might be attained through our own agency."*Nichomachean Ethics* [III, 1111b, 25].

theatre." More generally, for agents in strategic games to form intentions about arbitrary *outcomes* instead of only about their own strategies is for them to intend that they, *together with others*, act in a certain way. We shall see shortly that these intentions are at the heart of intention-based coordination in strategic games.

Intentions of the form "*I* intend that *we*..." introduce an explicit social or even cooperative aspect to strategic reasoning. They thus seems to spur the analysis towards more cooperative scenarios[6]. It is important to realise, however, that even if such intentions are carrying commitments that involve others, they are not binding *agreements* on others. Nothing precludes an agent from forming an intention that he cannot achieve by himself in a totally solipsistic manner, that is without agreeing with those involved in the "we" to play their part. Intentions of the form "I intend that we" are still individual intentions, even if they have a plural content. They are intentions that agents can form and hold *alone*. As such, they contrast with genuine we-intentions[7] and, to repeat, with genuine *agreements*. An agent cannot reach an agreement in isolation, but he can alone form an intention, the content of which involves action by others. Introducing such intentions, even though it brings some obvious social aspects into the analysis, does not turn non-cooperative scenarios into cooperative ones.

The achievement of such intention is of course seriously threatened if those included in the "we" do not intend to play their part. But this does not mean that the agent cannot form such an intention. This is highlighted by the fact that some authors, chiefly Bratman [1999, chap.8] and Velleman [1997], have put forward conditions under which an agent is *justified* in forming such an individual intention with a plural content. For them, an agent *should not*, even though he *can*, form such an intention without taking care of what the others intend. They argue that the agent must know that the others would also form the corresponding intention if they were to know that he has this intention. Conversely, they argue that an agent is not justified in forming an intention to do something together with others if he is not certain that learning about his intention is sufficient for the others to intend to play their part.

> Suppose now that the issue of whether *we* paint together is one that is obviously salient to both of us. [...] *I know* you would settle on this course of action if only *you were confident* about my appropriate attitude. I infer that if *you knew* that I intend that we paint, then you would intend that we paint, and we would then go on to paint together. Given this prediction, I form the intention that we paint [...]. [*idem*, p.155, my emphasis]

On this account, intentions of the form "I intend that we..." should be supported by a background knowledge of interdependent intentions. An agent is

[6] For more on cooperative games, see Myerson [1991, Sections 8-10].

[7] These are intentions of the form "we intend that ...". See Searle [1995] and Tuomela [1995].

justified in forming such an intention if he knows that the others would also adopt the same intention, if they knew that he so intends. In the context of strategic games, this means that an outcomes-intention that cannot be achieved by an agent alone is legitimate whenever its bearer knows that his co-players would also have this intention, if they knew he has it.

Mutual knowledge is the cornerstone of this account. In view of this, there is no need to restrict the analysis to actions-intentions in strategic games, as long as one complements it with an *epistemic* analysis. This is the route I take in this chapter. I incorporate outcomes-intentions in strategic games as in Definition 3.1.2, without imposing further constraints on the intention sets than those I imposed in Chapter 2. I then show how such intentions anchor coordination in strategic games, starting with the "easy" case of Hi-Lo games and then generalizing to arbitrary strategic contexts. Along the way I introduce epistemic models to capture the relevant knowledge conditions that are attached to outcome-intentions.

3.1.2. Definition. [Intentions in strategic games] Given a strategic game $\mathbb{G}$, an *intention set* $\iota_i \subseteq \mathcal{P}(X)$ for agent $i \in I$ is a set of sets of outcomes that is:

- *Internally consistent*: $\emptyset \notin \iota_i$ and $\iota_i \neq \emptyset$
- *Agglomerative*: If $A, B \in \iota_i$, then $A \cap B \in \iota_i$.
- *Closed under supersets*: If $A \in \iota_i$, and $A \subseteq B$ then $B \in \iota_i$.

The set $A \in \iota_i$ such that $A \subseteq B$ for all $B \in \iota_i$ is denoted $\downarrow\iota_i$. The intention set ι_i is said to be *generated* by $\downarrow\iota_i$. An *intention profile* ι is a vector of intention sets, one for each agent.

3.2 Coordination in Hi-Lo games

Hi-Lo games have become a benchmark for theories of inter-personal coordination. In these games the payoff structure counterbalances the uncertainty that usually hampers coordination. One coordination profile is obviously better for all players, and in experiments agents indeed massively choose it[8]. Standard game-theoretical arguments, however, are not able to pinpoint this profile as the only solution of Hi-Lo games. For that reason, most theories that claim to account for coordination start by showing that they can do it in the "easy" case of Hi-Lo games. As I show shortly, intention-based explanation indeed meets this benchmark.

[8] See Bacharach [2006, p.42-44] for references on experimental results. The presentation in this section is heavily based on Bacharach's extremely illuminating chapter on the "Hi-Lo paradox."

	Hi	Lo
Hi	2, 2	0, 0
Lo	0, 0	1, 1

Table 3.2: A Hi-Lo game.

Let me first introduce Hi-Lo games in more detail. They are a particular kind of coordination game, in which there is a *strictly Pareto-optimal* pure Nash equilibrium[9].

3.2.1. Definition. [Coordination Games] A *coordination game* is a strategic game $\mathbb{G}$ such that:

- $S_i = S_j$ for all $i, j \in I$
- $\pi(\sigma) \succ_i \pi(\sigma')$ for all σ such that $\sigma(i) = \sigma(j)$ for all $i, j \in I$ and σ' such that $\sigma'(i) \neq \sigma'(j)$ for some i and j.

Coordination games thus have a simple structure. Just as in the simple example of Table 3.1, one can view them as matrices where the "coordination profiles", the profiles where all agents play the same strategy, lie on the diagonal. As I mentioned, Hi-Lo games are coordination games where one coordination point, the $Hi - Hi$ profile in Table 3.2, is strictly Pareto-optimal[10].

3.2.2. Definition. [Weak and strict Pareto optimality] Given a strategic game $\mathbb{G}$, a strategy profile σ is *strictly Pareto-optimal* when $\pi(\sigma) \succ_i \pi(\sigma')$ for all agents $i \in I$ and profiles $\sigma' \neq \sigma$. It is *weakly* Pareto-optimal when $\pi(\sigma) \succeq_i \pi(\sigma')$.

3.2.3. Definition. [Hi-Lo games] A *Hi-Lo* game is a coordination game in which one of the profiles σ such that $\sigma(i) = \sigma(j)$ for all $i, j \in I$ is strictly Pareto-optimal.

The problem with Hi-Lo games is that no game-theoretic argument can single out the Pareto-optimal profile as the *only* rational solution. Agents cannot be sure, from game-theoretical reasoning alone, that their opponents will choose the strategy that leads to this profile. To see this, observe that all strategies might lead to coordination, and that all coordination points σ are pure Nash equilibria.

[9] See the Appendix to this chapter for the formal definition of Nash equilibrium and iterated removal of dominated strategies.

[10] The definition of Pareto optimality I use here comes from Colman and Bacharach [1997]. It is stronger than the standard game-theoretic notion. Myerson [1991, p.97], for instance, defines it as follows—he uses "outcomes" for what I here call "profiles": "An outcome of a game is (weakly) Pareto efficient iff no other outcome would make all players better off." One finds a similar definition in Osborne and Rubinstein [1994, p.7]. The definition of Pareto optimality I use states that no other outcome would make *any* player better off. I use this one because it makes it easier to draw the connection with Colman & Bacharach's work.

That is, for all i and $s_i \neq \sigma(i)$, we have that $\pi(\sigma) \succ_i \pi(s_i, \sigma_{j\neq i})$. This means that, for all agents, all strategies are compatible with playing a Nash equilibrium. What is more, no strategy is weakly dominated.

Here lies the whole "paradox" of Hi-Lo games. Despite strong intuitions that the *only* rational thing to choose in this game is Hi, and despite overwhelming empirical evidences that agents actually *do* choose Hi, standard game-theoretic arguments, in the words of Bacharach [2006, p.46], "fail to exclude" the sub-optimal profiles from the set of rationally plausible solutions.

To account for rational coordination in Hi-Lo games is thus to give an explanation of why the agents would choose Hi. There are many such explanations in the literature, but the details of most of them are rather tangential to my present concern. Instead of reviewing them, I briefly go over a very illuminating classification proposed by Bacharach [2006], in which we will be able to situate the intention-based account better.

To Bacharach, accounts of coordination in Hi-Lo games are either *re-specification* theories, *bounded rationality* theories or *revisionist* theories. The first type *re-describe* the Hi-Lo game in such a way that the Pareto-optimal profile becomes the only rational one, according to *standard* game-theoretical arguments. That is, re-specification theories try to show that the agents are in fact not facing the game as specified in Table 3.2, for instance, but another game in which Hi is the only rational choice. Along these lines, one can view coordination on Hi-Hi as the result of pre-play signals, as in [Aumann, 1987], or of repeated plays, as in [Aumann and Sorin, 1989].

Re-specification theories are often criticized because they, literally, play a different game. What they explain is not coordination in Hi-Lo *per se*, but rather in some other scenario. But, the argument goes, our intuition about the rationality of choosing Hi does not rest on any change of context. It seems as if choosing Hi is the only right thing to do, even in "pure" Hi-Lo games.

The accounts of the second type, bounded rationality theories, do stay within the limit of the original Hi-Lo story. To them "real" agents successfully coordinate because they do not reason as ideal game-theoretical agents would. A good example of such an alternative mode of reasoning is the *Stackelberg heuristic* of Colman and Bacharach [1997]. To them agents coordinate because they reason as if their opponents could read their mind. That is, they choose their strategy under the assumption that their opponents are able to anticipate this decision, *whatever* it is, and reply accordingly. If all agents reason this way one can show that they end up playing the Pareto optimal profile in Hi-Lo games.

I shall come back to this account in greater detail later, because it turns out to be closely related to the first intention-based account that I present. For now what is important is that, following Bacharach [2006], agents who reason this way are making an ungrounded assumption about their opponents' anticipation capacities. Indeed, they have no ground for believing that if they choose Hi their opponent will be able to anticipate this decision. As we saw, both Hi and

Lo are plausible choices in standard game-theoretical terms. As such, agents who follow the Stackelberg heuristic are not “fully” rational, as least in standard game-theoretical terms.

This type of bounded rationality account is also unsatisfying. In essence it argues that agents manage to coordinate because of some dubious or incomplete reasoning. But this, again, runs counter to a strong intuition about Hi-Lo games, namely that there is nothing wrong with choosing Hi. Quite the contrary, this seems like the only sensible thing to do, and *a fortiori* for agents who would reason correctly.

The third way of accounting for coordination keeps the Hi-Lo story intact, and does not look for reasoning mistakes or limitations. Rather, it tries to account for coordination by *revising* the very notion of rational choice. This is the approach championed, for instance, by Sugden [2003] and Bacharach [2006]. To them there are unavoidable “team” or “group” aspects to rational choice, and in Hi-Lo games they spur the agents toward the Hi-Hi solution.

The intention-based account that I explore in the following sections can also be seen as revisionist. Instead of invoking teams or groups, it rests on the idea that rational decision making for planning agents should take previously adopted intentions into account. That is, a rational choice in a strategic game with intention is not only one that is rational in the classical game-theoretical sense, but also one in which the agents follow the intentions they might have formed before playing the game, and in which they take into account what they know about each others’ intentions.

3.3 Intentions and rational expectations

We saw in the previous chapter that intentions can break ties between equally desirable options, provided they are payoff-compatible. This can already be seen as a vector for coordination, at the personal level. It provides agents with an additional criterion to discriminate future courses of action. They can better anticipate their own choices and make further decisions on that basis. By generalizing the notion of payoff-compatibility to contexts of strategic interactions, this tie-breaking effect turns into an anchor for *inter*-personal coordination, at least in Hi-Lo games.

One has, however, to be careful in defining what payoff-compatible intentions are in strategic interaction. If we directly use the single-agent version of this requirement, intention-based reasoning quickly runs up against standard game theoretical rationality.

Consider, for example, the game in Table 3.3. Agents who would have payoff-compatible intentions, in the sense of Chapter 2, would be at odds with basic game-theoretic assumptions. Recall that for an intention set to be payoff-compatible is the same as stating that its smallest element $\downarrow\iota_i$ has to contain only

	A	B
a	0, 0	0, 7
b	7, 0	1, 1

Table 3.3: A game where parametric payoff-compatible intentions go against standard game-theoretical reasoning.

outcomes that are most preferred. Since each agent has a unique most preferred outcome in this game—here I directly take the profiles as outcomes—there is only one intention set per agent that is payoff-compatible, namely the one generated by $\{(b, A)\}$ for 1 and the one generated by $\{(a, B)\}$ for 2. But observe that a is a strictly dominated strategy for the first agent, and so is A for the second. This means that, to achieve his intention, each agent needs the other to play a strictly dominated strategy[11].

This should not come as a surprise. Payoff-compatibility of intentions is tailored for single-agent contexts, where one does not base his decision on what he expects other rational agents to do. But if we want an account of coordination that does not fall into the "bounded rationality" category, in which the agents' reasoning violates standard game-theoretical assumptions, we have to adjust intention-based reasoning to *mutual expectations*[12]. Rational choice in strategic games is, in other words, a matter of maximizing expected payoffs given what the agents expect each other to do. Any account of coordination that builds on the traditional standard of rationality in games has to take these mutual expectations into account.

In the parametric setting we had a relatively uncontroversial criterion for rational choice, namely maximization of expected utility. In interactive contexts the mutual dependencies of expectations has yielded a whole array of solution concepts. For instance, Nash equilibrium and strong dominance each isolate a different set Γ of rationally plausible strategy profiles. To accommodate this plurality I make payoff-compatibility relative to solution concepts.

3.3.1. Definition. [Feasible outcomes] Given a solution concept $\Gamma \subseteq \Pi_{i \in I} S_i$, an outcome x is said to be *feasible according to* Γ, or Γ*-feasible* iff there is a profile $\sigma \in \Gamma$ such that $\pi(\sigma) = x$.

3.3.2. Definition. [Payoff-compatible intentions - the general case] Let A^* and B^* denote the sets consisting of all Γ-feasible elements of A and B, respectively.

[11] This is even more problematic given that these intentions are of the form "I intend that we..."—i.e. that the agents cannot realize them unilaterally. But this problem concerns the theory of intentions more than the clash between these intentions and game-theoretic rationality. I shall return to more intention-based criteria in a moment.

[12] Recall the remarks about mutual expectations in the Introduction (Section 1.1.2).

An intention set ι_i is said to be *payoff-compatible* whenever $A \in \iota_i$ if $A \cup B \in \iota_i$, $A^* \neq \emptyset$ and $x \succeq_i y$ for all $x \in A^*$ and $y \in B^*$.

It is easy to check that if we consider "games" where there is only one agent and where Γ is the set of most preferred outcomes, this new condition boils down to the one I introduced in the previous chapter (Definition 2.5.5). Recall that the agent was said to have payoff-compatible intentions when, given the fact that he does not intend B, all the elements of which are strictly better than all the elements of another set of outcomes A, we could conclude that he does not intend $A \cup B$ either. The idea here is essentially the same, except that it is adapted to rational expectation. An agent has payoff-compatible intentions when, given the fact that he does not intend B, all the *feasible* elements of which are strictly better than all the *feasible* elements, if any, of another set of outcomes A, we can conclude that he does not intend $A \cup B$ either. In other words, if all the outcomes that the agent can expect in a set B are strictly better than all the outcomes he can expect in a set A, and yet he does not intend B, then he does not intend $A \cup B$ either. Not surprisingly, we get a characterization of the payoff-compatible intentions that is analogous to the one we saw in the previous chapter.

3.3.3. Fact. Let Γ be a solution concept and $\Gamma_{\succeq_i}$ be the set of most preferred Γ-feasible outcomes for agent i. For all i and ι_i as in Definition 3.1.2, ι_i is payoff-compatible if and only if there is a non-empty $A \subseteq \Gamma_{\succeq_i}$ such that $\iota_i = \{B | \; A \subseteq B\}$.

Proof. Essentially the same as for Fact 2.2.9, adapted to feasible sets. ■

Generalized payoff-compatibility thus ensures that agents intend to realize some of their most preferred *feasible* outcomes. That is, if an agent is not indifferent between the outcomes he can rationally expect, his intentions "pick" some of those he prefers most. This is exactly what happens in Hi-Lo games.

3.3.4. Corollary (Intention overlap in Hi-Lo games). *For any agent i in a Hi-Lo game,* $\downarrow\iota_i = \{\pi(\sigma^*)\}$*, where* σ^* *is the strictly Pareto-optimal profile.*

Agents with payoff-compatible intentions are thus bound to agree on what they intend to realize in Hi-Lo games. But this only explains why planning agents *intend* to realize the outcome of this profile, and not why they would *actually* coordinate. To coordinate successfully on the basis of such overlapping intentions they still need to *act on them.* That is, these intentions must somehow translate into action or, in the words of Bratman [1987] be "conduct controlling." Observe, furthermore, that both these intentions are of the form "I intend that *we* achieve the Pareto-optimal solution." As I mentioned at the end of Section 3.1, there are arguably cases where the agents are *not* justified in having such intentions, because they lack some required information about the intentions of others. The use of payoff-compatible intentions thus requires one to spell out more explicitly the volitive commitment that comes with intentions and the epistemic conditions under which Hi-Lo games are played.

3.4 Epistemic models for games with intentions

At least since the work of Harsanyi [1967-68] and Aumann [1976], *epistemic models* have been used within the epistemic programme in game theory to understand how rational agents base their choices on what they know and believe about their opponents' preferences, rationality and expectations[13]. The epistemic characterization of the elimination of strongly dominated strategies is a classical example of what can be shown with these epistemic models. Brandenburger and Denkel [1987] showed that if, in an epistemic model, all agents are rational and commonly believe that all others are rational, then they do not play a strategy that is strictly dominated. In other words, rationality and common belief in rationality are *sufficient conditions* for agents to choose strategies that are not strictly dominated. In this section I follow a similar line: by building epistemic models for games with intentions, I spell out sufficient conditions for coordination with payoff-compatible intentions, and at the same time make more explicit the background knowledge that supports individual intentions with a "we" content.

An epistemic model of a given game $\mathbb{G}$ is a structure that represents what the agents might know, believe and prefer in diverse scenarios or plays of that game. Two main types of models have been used in the literature: *type spaces* and the so-called *Aumann*- or *Kripke-structures*. Both represent the possible plays of the game as *states*, where each agent chooses a particular strategy and has information about the others. Type spaces and Aumann structures differ in the way they represent this information. The first represent the agent's information probabilistically, while the second use partitions or "accessibility relations". As noticed by Brandenburger [2007], these two modelling paradigms have given rise to different styles of epistemic analysis. The probabilistic nature of type spaces have naturally led towards *belief*-based characterizations. Aumann or Kripke structures, on the other hand, have mostly provided *knowledge*-based characterizations[14]. In what follows I use the latter, and provide a knowledge-based analysis in which the conditions for "I intend that we" are easily spelled out.

Let me first give the formal definition of an epistemic model.

3.4.1. Definition. An *epistemic model* $\mathbb{M}$ of the game $\mathbb{G}$ is a tuple $\langle W, f, \{\sim_i\}_{\in I}\rangle$ such that:

- W is a set of states.

- Let $\mathcal{F}(X)$ be the set of all filters over the set of outcomes X in $\mathbb{G}$. Then $f : W \to \Pi_{i\in I} S_i \times \Pi_{i \in I}\mathcal{F}(X)$ is a function that assigns to each $w \in W$ a pair (σ, ι) of strategy and intention profile. From convenience I write $\sigma(w)$

[13] For references see the footnote on page 1.

[14] The work of Baltag and Smets [Unpublished manuscript] and Mihalache [2007] are notable exceptions.

and $\iota(w)$ for the σ (alternatively the ι) such that $f(w) = (\sigma, \iota)$, and $f_i(w)$, $\sigma_i(w)$ and $\iota_i(w)$ for the i^{th} component of these (pairs or) profiles.

- $\sim_i$ in an equivalence relation on W such that if $w \sim_i w'$ then $f_i(w) = f_i(w')$. I write $[w]_i$ for $\{w' : w \sim_i w'\}$.

A *pointed model* $\mathbb{M}, w$ is an epistemic model for the game $\mathbb{G}$ together with a distinguished state w, the *actual play* of $\mathbb{G}$.

This is essentially an extension to strategic games *with intentions* of the models proposed by Aumann [1994]. As I wrote above, each state w represents a possible play of the game. At each of them the agents are making a particular strategy choice, $\sigma_i(w)$, and have some intentions, $\iota_i(w)$. A set of states $E \subseteq W$ is called an *event*. It is the set of states where the event E takes place.

The information of each agent is represented as in Kripke models for epistemic logic[15]. The accessibility relation $\sim_i$ connects a state w to all the states that i cannot distinguish from it or, in other words, to all the states that i considers possible at w. As just mentioned, I use here a "knowledge-based" representation, which boils down to assuming that information is veridical, i.e. agents always consider that the current state is possible, and strongly introspective, i.e. agents are always aware of what they consider possible and what they do not. In technical terms, this means that $\sim_i$ is an equivalence relation, i.e. that it is reflexive, transitive and symmetric.

In these models it is generally assumed that i knows that E at w whenever E takes place in all states w' that i considers possible, i.e. whenever $[w]_i \subseteq E$. To paraphrase Cozic [2005, p.290], to know that an event takes place in these models is to exclude that it might not take place. Following common practice in the literature, I denote by $K_i(E)$ the set of states where i knows that E takes place. An event E is considered possible at a state w whenever there is a w' that i considers possible at w in which E takes place.

With this in hand, one can see more clearly the conditions that are imposed on $\sim_i$. As already mentioned, reflexivity, transitivity and symmetry of this relation make information in game models veridical and strongly introspective. Reflexivity ensures truthfulness: if i knows that E at some state then E takes place at that state. In formal terms, $K_i(E) \subseteq E$. Transitivity ensures "positive" introspection: whenever an agent knows that E takes place he also knows that he knows. Symmetry ensures "negative" introspection: if an agent does not know that E takes place he at least knows that he does not know[16].

These are classical assumptions that make K a "knowledge" operator. In the literature as well as in the above definition, it is also assumed that agents know their strategy choice at each state. Similarly, I assume that agents know their

[15] See the references in the footnote on page 2.

[16] In our models these two conditions boil down to $K_i(K_i(E)) \subseteq K_i(E)$, and $K_i(W - K_iE) \subseteq W - K_iE$, where $W - A$ is the complement of A in W.

own intentions. If we take I_iA to be the set of states where A is in the intention set of i and s_i the set of states where i chooses s_i, that is $I_iA = \{w : A \in \iota_i(w)\}$ and $s_i = \{w : \sigma_i(w) = s_i\}$, one can check that the condition "$f_i(w) = f_i(w')$ if $w \sim_i w'$" ensures that $K_i(I_iA) \subseteq I_iA$ and $K_i(s_i) \subseteq s_i$. That is, at all the states w' that the agent i considers possible at a state w, he plays the same strategy ($\sigma_i(w) = \sigma_i(w')$) and has the same intentions ($\iota_i(w) = \iota_i(w')$)[17]. What i might be uncertain about is the strategy choices and intentions of the other players. In cases where $w' \sim_i w$ if *and only if* $\sigma_i(w) = \sigma_i(w')$, for instance, he considers possible all combinations of actions of the other agents. But he might be better informed about the current state, and thus not consider all choices of others possible. Agent i might know, for instance, that j does not play strictly dominated strategies, and thus that i does not consider the state w' possible because j plays such a strategy in that state.

It might be helpful at this point to look at an example. Consider again the Hi-Lo game of Table 3.2, and assume that the set of outcomes is the set of profiles itself. One of its possible models is depicted in Figure 3.1. It has four states, which are in one-to-one correspondence with the strategy profiles of the game. The players are as uninformed as they can be. At each state, they consider all choices of their opponent possible. Agent 1, for example, considers at $Hi - Hi$ that 2 might play Hi as well as Lo. But 1 knows what he plays at $Hi - Hi$. In all states that he considers possible, he plays Hi.

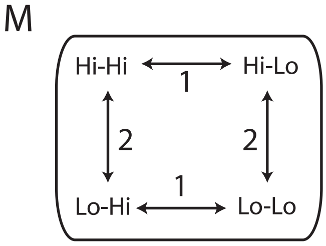

Figure 3.1: Epistemic model for the Hi-Lo game of Table 3.2. The arrows represent the relation $\sim_i$ for each player.

This model can be completed by many different assignments of intentions to the states. It might be, for example, that at $Hi - Hi$ both agents have payoff-compatible intentions. As we saw in the last section, for Γ the pure Nash equilibrium solution concept, this means that $\iota_i(Hi - Hi) = \{Hi - Hi\}$ for both agents. Given that the agents know what they intend, that would mean that

[17] These conditions are illustrated in Figures 5.1 and 5.2, on page 94 and 95.

$\iota_1(Hi - Lo)$ must also be $\{Hi - Hi\}$, and the same for 2 at $Lo - Hi$. But 1 could very well have a different intention at this state. He might play Lo at $Lo - Hi$ because he in fact intends $Lo - Lo$, i.e. $\iota_1(Lo - Hi) = \{Lo - Lo\}$. That would mean that at $Hi - Hi$ agent 2 is uncertain of 1's intention. As far as he knows, 1 might intend $Hi - Hi$ as well as $Lo - Lo$.

It is worth stressing that this is just *one* possible completion of the set of states Figure 3.1. There might be other models with more states, such as the one on page 50, or other models with the same number of states but different assignments of intentions and strategy at each states. It might be, for instance, that in Figure 3.1 agent 1's intentions are payoff-compatible in *all* states. That is, it might be that $\iota_1(w) = \{Hi - Hi\}$ for all states $w \in W$. In this case agent 2 knows agent 1's intentions at $Hi - Hi$. In all states that 2 considers possible, 1 intends to achieve the Pareto-optimal profile.

This second completion of the set of states Figure 3.1 features a notable discrepancy between what 1 intends, chooses and knows at $Lo - Hi$. At this state he plays Lo even though he intends to achieve $Hi - Hi$. What is more, he does not even consider it possible to achieve this intention. At $Lo - Hi$ he does not, in a very strong sense, act on his intention to achieve $Hi - Hi$.

As I hinted at the end of Section 3.3, this idea of "acting on one's own intention" is one of the key ingredients of an intention-based account of coordination. Thanks to epistemic models, it can be made precise.

3.4.2. Definition. [Intention-Rationality] A player i is said to be *intention rational* at a pointed model $\mathbb{M}, w$ if and only if and

$$\pi(\sigma_i(w)) \cap {\downarrow}\iota_i(w) \neq \emptyset$$

The set of states where i is intention rational is noted IR_i. Formally, $IR_i = \{w : i$ is intention-rational at $\mathbb{M}, w\}$

An agent is thus intention-rational at a state when he chooses an action by which he can achieve at least one outcome he intends. Put as a contrapositive, what this means is that an agent is intention-irrational at a state w when he excludes by his own decision the achievement of his intentions. In other words, an agent is intention-irrational when he is not doing anything to achieve what he intends.

3.5 Coordination with payoff-compatible intentions

We already know that if all agents have payoff-compatible intentions in Hi-Lo games, then their intentions will "overlap" on the Pareto-optimal profile. If, furthermore, each agent is intention-rational, that is if they act on these intentions, then they successfully coordinate.

3.5.1. Fact. For any Hi-Lo game the following holds:

1. For any of its pointed models $\mathbb{M}, w$, if both agents are intention-rational, have payoff-compatible intentions and Γ is the pure Nash equilibrium solution concept then $\sigma(w)$ is the Pareto-optimal strategy profile σ^* of that game.

2. If σ^* is the Pareto-optimal strategy profile of that game, then we can construct a pointed model $\mathbb{M}, w$ of such that $\sigma(w) = \sigma^*$, all agents are intention-rational and their intention are payoff-compatible.

Proof.

1. Let x be $\pi(\sigma^*)$. For any agent i, we know from Fact 3.3.3 that, because he has payoff-compatible intentions, $\downarrow\iota_i(w) = \{x\}$. Now, because i is also intention-rational, we also know that $\pi(\sigma_i(w)) \cap \downarrow\iota_i \neq \emptyset$, which is just to say that $x \in \pi(\sigma_i(w))$, which means that there is a σ' such that $\sigma'(i) = \sigma_i(w)$ and $\pi(\sigma') = x$. But observe that by the very definition of Hi-Lo games, there can be no other σ' such that $\pi(\sigma') = x$. This means that σ' can only be σ^*, and so we conclude that $\sigma_i(w) = \sigma^*(i)$. Since we took an arbitrary i, this is also the case for all i, and thus $\sigma(w) = \sigma^*$.

2. Just fix $\sigma(w) = \sigma^*$ and $\downarrow\iota_i(w) = \{\pi(\sigma^*)\}$, for all i.

■

Part 1 of this result is our first intention-based account of coordination. Intention-rationality and payoff-compatible intentions are sufficient for coordination in Hi-Lo games. The second part of the result means that one can always look at coordination on the Pareto-optimal profile from an intention-based perspective. That is, we can always model the agents *as if* they coordinate on the basis of their intentions[18].

It is important to appreciate that this result is *not* epistemically loaded. It can be that agents successfully coordinate even though they consider it possible that the others will not enact their intentions or that they do not have payoff-consistent intentions.

[18] The "as if" is important here. One can always construct epistemic models for Hi-Lo games in which the agents coordinate on the Pareto-optimal profile against all odds, so to speak. At a state w where $\sigma(w)$ is the Pareto-optimal profile, it can very well be that none of the agents are intention-rational or have payoff-compatible intentions, and that the relations $[w]_i$ are such that the agents are completely uncertain about what the others do and intend. Epistemic characterizations, even of standard solution concepts, cannot rule out such cases. Coordination can just happen from sheer luck, after all. To draw a parallel with decision theory, it might well be that the decision maker's choices happen to maximize expected value, even if his overall choice behaviour is not representable by a payoff function on outcomes. The "as if" in a game-theoretic epistemic characterization, as in a decision-theoretic representation, just means that when coordination occurs there is a possible explanation for it using the condition stated in the result.

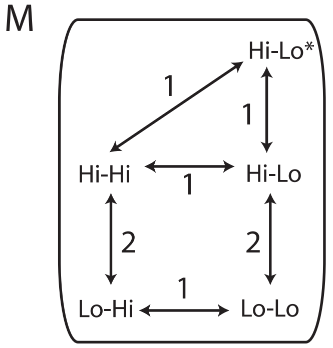

Figure 3.2: Another epistemic model for the Hi-Lo game of Table 3.2.

Look for example at the state $Hi - Hi$ in the model of Figure 3.2. Assume that at every state the agents have payoff-consistent intentions, except at the additional $Hi - Lo^*$ state where 2 intends $Lo - Lo$. At $Hi - Hi$, agent 1 has doubts about 2's intentions. As far as he knows, 2 might as well be intention-irrational or have payoff-incompatible intentions.

Fact 3.5.1 thus leaves plenty of room for cases of coordination in Hi-Lo where the agents are uncertain about the others' intentions. In a way, this shows how "easy" it is to coordinate in these games. The Pareto-optimal profile leaves no room for agents with payoff-compatible intentions to intend anything else. This is indeed not particular to Hi-Lo games. Intention overlap in the context of payoff-compatibility is closely related to the existence of such an outcome.

3.5.2. FACT. [Coordination and Weak Pareto-optimality] For any strategic game $\mathbb{G}$ the following are equivalent.

1. There is a weakly Pareto-optimal profile σ^*.
2. There is an epistemic pointed model $\mathbb{M}, w$ for $\mathbb{G}$ such that at w all agents have payoff-compatible intentions, taking $\Gamma = \Pi_{i \in I} S_i$, and $\bigcap_i \downarrow\iota_i \neq \emptyset$.

Proof. From (2) to (1). Take such an epistemic model and look at any $x \in \bigcap_i \downarrow\iota_i$. Assuming that $\Gamma = \Pi_{i \in I} S_i$, Fact 3.3.2 directly gives us that there is a profile σ such that $\pi(\sigma) = x$ and that for all i and all profile σ', $x \succeq_i \pi(\sigma')$.

From (1) to (2), take any such $\pi(\sigma^*)$. Any pointed model $\mathbb{M}, w$ built according to Definition 5.2.1 in which $\downarrow\iota_i = \{\pi(\sigma^*)\}$ for all i at w will do. ■

This result shows that payoff-compatible intentions are especially suited to drive coordination in games where the agents' preferences converge towards a most preferred outcome. But the reader should also appreciate that in most cases agents will not be able to achieve these intentions by themselves. That is, whenever the set of unanimously preferred outcomes is “small enough”, payoff-compatible intentions turn into intentions of the form “I intend that *we* reach the outcome that we all most prefer.”

This kind of intention, as I mentioned at the end of Section 3.1, should arguably be supported by some information about the intentions of others. Namely, an agent who has intentions of the form “I intend that we. . . ” should *know* that if the others—those included in the we—*knew* he has this intention, they would also go on and adopt the same intention. Given that intentions are generally taken as conduct controlling, I will assume that this last clause means “adopt the same intention *and act on it*.” The formal counterpart of this condition is thus the following.

3.5.3. Definition. [Epistemic support] The intention of i to achieve A at w is said to be *epistemically supported* whenever, for all $j \neq i$, and all $w' \sim_i w$, if $w'' \in (IR_i \cap I_iA)$ for all $w'' \sim_j w'$, then $w' \in IR_j \cap I_jA$.

The reader can check that this actually corresponds to the fact that i knows that if j knows that i is intention-rational and intends to achieve A, then j is also intention-rational and intends to achieve A.

As might be suspected, at $Hi - Hi$ in the model of Figure 3.2 the intention of 1 to achieve $Hi - Hi$ is *not* epistemically supported. Indeed, 2 knows at $Hi - Lo^*$ both that 1 is intention-rational and that 1 intends to achieve $Hi - Hi$. He (2) does not, however, intend $Hi - Hi$. This means that, at $Hi - Hi$, agent 1 considers it possible that 2 will still not play his part in achieving $Hi - Hi$ even though 2 recognizes that this is what 1 intends.

Epistemic support is thus not necessary for successful coordination in Hi-Lo games. Because of its conditional content, it is not sufficient either. Look for example at the set of states of Figure 3.1, completed with the intentions specified in Table 3.4. At both $Hi - Hi$ and $Hi - Lo$ agent 2 does not know whether 1 intends $Hi - Hi$. But this means that at $Hi - Lo$, in all states that 1 considers possible the implication “if 2 knows that 1’s is intention-rational and intends $Hi - Hi$ then 2 also intends $Hi - Hi$” is trivially true. In other words, at $Hi - Lo$ agent 1’s intention to achieve $Hi - Hi$ is epistemically supported. A similar argument, this time because 2 is not intention-rational at $Hi - Lo$, shows that 2’s intention to achieve $Hi - Hi$ is also epistemically supported. Both agents thus intend $Hi - Hi$ with the required epistemic support, and yet at $Hi - Lo$ they fail to coordinate.

Of course, one could object that in such a case the agents do not have a “genuine” epistemic support for their intention to $Hi - Hi$. In no state that they consider possible is the antecedent of the epistemic support condition met. To avoid such cases one can strengthen this condition.

3.5.4. Definition. [Strong epistemic support] The intention of i to achieve A at w is *strongly epistemically supported* whenever, for all $j \neq i$, $w' \in IR_j \cap I_jA$ for all $w' \sim_i w$ and $w'' \in (IR_i \cap I_iA)$ for all $w'' \sim_j w'$.

State	$\downarrow\iota_1(w)$	$\downarrow\iota_2(w)$
$Hi-Hi$	$Hi-Hi$	$Hi-Hi$
$Hi-Lo$	$Hi-Hi$	$Hi-Hi$
$Lo-Hi$	$Lo-Lo$	$Hi-Hi$
$Lo-Lo$	$Lo-Lo$	$Lo-Lo$

Table 3.4: The intentions for the model in Figure 3.1.

In two-agents cases this boils down to saying that 1 knows that 2 knows that 1 is intention-rational and intends to achieve A, *and* that 2 has the corresponding intention. Strongly epistemically supported intentions that $Hi-Hi$ *are* sufficient for successful coordination.

3.5.5. Fact. [Second account of intention-based coordination] For any Hi-Lo game the following holds:

1. For any of its pointed models $\mathbb{M}, w$, if all agents have strongly epistemically supported intentions to achieve $\{\pi(\sigma^*)\}$, then $\sigma(w)$ is the Pareto-optimal strategy profile σ^* of that game.

2. If σ^* is the Pareto-optimal profile of that game, then we can construct a pointed model $\mathbb{M}, w$ such that $\sigma(w) = \sigma^*$ and all agents have strongly epistemically supported intentions that $\{\pi(\sigma^*)\}$.

Proof. The second part follows the same step as in the proof of fact 3.5.1. For the first part, observe that it follows directly from $K_i(E) \subseteq E$ that, at any state w where both agents have strongly epistemically supported intentions to achieve $\{\sigma^*\}$, they are intention-rational and their most precise intention is $\{\sigma^*\}$. This, we know from Fact 3.5.1, ensures that $\sigma(w) = \sigma^*$. ■

This result rests essentially on the fact that knowledge is veridical in models for games with intentions. If 1 knows that 2 knows that 1 intends $Hi - Hi$ and is intention-rational, then 1 *does* so intend. In fact, one can bypass this embedded knowledge condition. Mutual knowledge of intention-rationality and payoff-compatibility is also sufficient for coordination.

3.5.6. Fact. [Third account of intention-based coordination] Let IPC_i be the set of states w of an epistemic model for games with intentions where $\iota_i(w)$ is payoff-compatible, and Γ be the set of pure Nash equilibria. Then for any Hi-Lo game the following holds:

1. For any of its pointed model $\mathbb{M}, w$, if $w \in K_i(IR_j \cap IPC_j)$ for all $i, j \in I$ then $\sigma(w)$ is the Pareto-optimal strategy profile σ^* of that game.

2. If σ^* is the Pareto-optimal profile of that game, then we can construct a pointed model $\mathbb{M}, w$ such that $\sigma(w) = \sigma^*$ and all agents have strongly epistemically supported intentions that $\{\pi(\sigma^*)\}$.

Proof. Again, the second part is obvious and the first is a direct consequence of $K_i(E) \subseteq E$ and Fact 3.5.1. ■

This characterization of coordination situates more explicitly the accounts based on payoff-compatible intentions with respect to standard game-theoretical reasoning in strategic games. The Hi-Hi profile is a pure Nash equilibrium of that game. Such profiles have been characterized by Aumann and Brandenburger [1995] in terms of rationality and mutual knowledge of strategy choice. More precisely, they have shown that at any pointed model $\mathbb{M}, w$ for a strategic game with two players, if both are rational and know the strategy choice of the other, then they play a Nash equilibrium at w. These two conditions are more or less explicitly at work in Fact 3.5.6. First, all agents can "deduce" the other's strategy choice, and thus meet Aumann and Brandenburger's mutual knowledge requirement, from the fact that they know that the others are intention-rational and have payoff-compatible intentions[19]. Second, the notion of feasibility, built-in in payoff-compatibility, secures the rationality requirement[20]. Recall that, in Section 3.3, I introduced the idea of a feasible outcome precisely to keep the intention-based account within the bounds of standard game-theoretical reasoning or rational expectations. Payoff-compatibility of intentions just gives the extra push for the agents to go beyond these standard assumptions, and by the same token to ensure coordination.

In other words, with this third characterization of coordination we can see better how the intention-based account, with payoff-compatible intentions, falls into the *revisionist* category. It preserves, on the one hand, the Hi-Lo game scenario and respects standard game-theoretical reasoning. What it provides, on the other hand, is a new criterion for rational decision, one which takes seriously the ability of planning agents to form intentions and with it the volitive commitment that these intentions carry. It supplements the traditional notion of instrumental rationality with considerations regarding intentions in interactive contexts.

This intention-based account is of course "socially" oriented, in comparison with more purely competitive ones like the Stackelberg heuristic which I present in the next section. In two of the characterizations of coordination, the agents

[19] This, it should be stressed, is a direct consequence of the fact that there is only one most preferred feasible outcome in these games. In general, knowing that an agent is intention-rational and hasve payoff-compatible intentions is *not* enough for another agent to know which strategy the first plays. This third account of coordination thus show how agents can combine their knowledge of the others' intentions with their knowledge of the structure of the game to make their decision.

[20] It is not essential for now to go into details of what "rationality" means in Aumann & Brandenburger's characterization. I come back to it in Chapter 5.

explicitly take the intentions of others into account. But, as we shall see in Section 3.7, reasoning with payoff-compatible intentions remains a broadly individualistic process. The intention-based account of coordination in Hi-Lo games is thus not fully cooperative, but nor is it purely competitive either.

3.6 Stackelberg heuristic and intention-based coordination

Intention-based coordination with payoff compatibility also provides an alternative to another account of coordination, which is not revisionist but rather a *bounded rationality* account. This account, the Stackelberg heuristic, is a mode of strategic reasoning proposed by Colman and Bacharach [1997]. The basic idea is that players reason as if their deliberation was "transparent" to the others. That is, they make their decisions under the assumption that, whatever they decide, the others will be able to anticipate their decisions and react accordingly.

It is important to realize that this assumption is much stronger than standard game-theoretic ones. Recall from the previous chapter that ideal game-theoretical agents are assumed to be "intelligent", in the sense that they can reach any conclusion the modeller is able to reach. If, for example, we conclude that agent i will not play his strategy s_i because it is strictly dominated, then all agents in the game can also reach this conclusion and react accordingly. But in many games, as in Hi-Lo, the agents are not able to anticipate the others' choices with game-theoretical reasoning alone. If an agent chooses Hi he must do so for reasons that are not, strictly speaking, game-theoretical. The assumption behind the Stackelberg heuristic is that other agents can "see" this reason, *whatever it is*. Paraphrasing Colman and Bacharach [1997, p.13], the agents reason as if the others can read their mind, and react accordingly. Another way to look at the Stackelberg heuristic is that agents reason as if they were not, in fact, playing a strategic game. Rather, they think of themselves as moving first in an extensive game with perfect information[21]. The others, they think, can witness this move and reply accordingly. Put that way, the principle behind the Stackelberg heuristic indeed looks like what Bacharach [2006, p.50-51] calls "magical thinking". On game-theoretical grounds alone, the players are absolutely not justified in reasoning that way. They have no reason to think that the others can anticipate all their decisions.

Formally, to define the *Stackelberg solution* of a strategic game $\mathbb{G}$ with two agents we need a few preliminary notions.

3.6.1. Definition. [Best response and Stackelberg Payoffs]

[21] The reference to H. F. von Stackelberg (1905-1946) comes from this idea. A *Stackelberg model* is a model where two firms choose sequentially the output price of some good. See Osborne [2004, p.187-189] for more details.

- Given a strategy s_i of i in a strategic game $\mathbb{G}$ with two agents, a *best response* for j, noted $\beta_j(s_i)$, is a strategy s_j such that for all s'_j, $\pi(s_j, s_i) \succeq_j \pi(s'_j, s_i)$.

- A *Stackelberg outcome* $h_1(s_1)$ of player 1 from the strategy s_1 a x such that $x = \pi(s_1, \beta_2(s_1))$, and similarly for player 2: $h_2(s_2)$ is any x such that $x = \pi(\beta_1(s_2), s_2)$.

In other words, a Stackelberg outcome of agent i's strategy s_i is an outcome he would get if his opponent were to play a best response against s_i. This is indeed what would happen if i were to choose s_i first in an extensive game of perfect information. The *Stackelberg solution* of a game, if it exists, is a profile where both players achieve a most preferred Stackelberg outcome.

3.6.2. Definition. [Stackelberg solubility and Stackelberg solutions] Let s_i^h be a strategy of i that yields a most preferred Stackelberg outcome. A two-agent strategic game $\mathbb{G}$ is *Stackelberg soluble* if there is a σ^h such that $\sigma^h(i) = s_i^h$ for all $i \in I$. σ^h is called a *Stackelberg solution* of that game.

It is easy to see that whenever a Stackelberg solution exists it is a Nash equilibrium. Colman and Bacharach have restricted their analysis to strategic games where, for both players $i \in I$, there is a unique s_i^h. Under this assumption, the next fact follows straightforwardly.

3.6.3. Fact. [Colman and Bacharach, 1997] In every 2 player game with more than one Nash equilibrium, σ is the Stackelberg solution iff it is the strictly Pareto-optimal outcome.

As a direct corollary we obtain that $Hi - Hi$ is also the Stackelberg solution in two-player Hi-Lo games, and so that the Stackelberg heuristic accounts for coordination in these contexts.

This result rests heavily on the simple structure of Hi-Lo games. Just as with payoff-compatible intentions, if all agents reason with the Stackelberg heuristic the Pareto-optimal profile is all that is left for them to choose. This similarity is not a coincidence. In games like Hi-Lo, where there is a unique most preferred Stackelberg outcome for each agent, the existence of a Stackelberg solution ensures overlap of intentions on it, and vice-versa.

3.6.4. Fact. The following are equivalent for any two-agent strategic game $\mathbb{G}$ with intentions in which s_i^h is unique, ι_i is payoff-compatible for both agents, and Γ is the set of pure Nash equilibria.

- $\mathbb{G}$ is Stackelberg soluble with σ^h.

- $\bigcap_{i \in I} \downarrow\iota_i = \{\pi(\sigma^h)\}$.

Proof. $\mathbb{G}$ is Stackelberg soluble with σ^h iff σ^h is a Nash equilibrium, thus iff $\pi(\sigma^h)$ is a Γ-feasible outcome. Now, observe that because s_i^h is unique for both i there can be no other Nash equilibrium $\sigma' = (s_1', s_2')$ such that $\pi(\sigma') \succeq_i \pi(\sigma^h)$. Indeed, if there were such a σ', in virtue of it being a pure Nash equilibrium we would have $(s_1', \beta(s_1')) = (\beta(s_2'), s_2')$. But this would contradict our assumption, since s_1^h is the *unique* strategy that yields the most preferred Stackelberg outcome, and similarly for player 2. This means that $\mathbb{G}$ is Stackelberg soluble with σ^h iff $\pi(\sigma^h)$ is the unique most preferred feasible outcome, both for 1 and 2. This, in turn, by Fact 3.3.3, happens iff $\iota_i = \{\pi(\sigma^h)\}$ for both players. ■

The Stackelberg heuristic and the payoff-compatibility condition thus yield the same conclusions about what the agents should choose or intend in games like Hi-Lo. But they are doing so on different bases. In the case of Stackelberg reasoning the agents reason under the assumption that the others will anticipate their choices. Even though it serves well in Hi-Lo games, this assumption is nevertheless not grounded in any game-theoretic reasons. To repeat, agents reasoning with the Stackelberg heuristic are not “fully” rational. They are making assumptions that ideal game-theoretic agents would not make. With payoff-compatible intentions, on the other hand, the agents are fully *intention*-rational, not in the technical sense of Definition 3.4.2, but in the sense that they are taking seriously their capacity to form an intention. Here, this capacity is constrained by a policy of intending the outcomes that they prefer the most among those they can rationally expect. The result above thus shows that, in games where there is a unique most preferred Stackelberg outcome for each agent, one can account for coordination by revising in a natural way what “rational” means, instead of attributing ungrounded assumptions to the agents.

3.7 Limits of payoff-compatibility of intentions

The Stackelberg heuristic and the three accounts of coordination that I presented in Section 3.5 rest heavily on the simple structure of Hi-Lo games. As we saw in Fact 3.5.2, payoff-compatibility of intention ensures coordination when there is a weakly Pareto-optimal profile. Fact 3.6.4 tells us that is no different for the Stackelberg heuristic.

This dependence on the existence of Pareto-optimal profiles is a double-edged sword for both accounts. It provides, on the one hand, a simple explanation of coordination in games like Hi-Lo. But in games where the preferences of the agents over the set of feasible outcomes diverge, payoff-compatible intention can lead the agents out of the feasible set. Look for example at the *Battle of the Sexes*, displayed in Figure 3.5. Here two agents, whom Luce and Raiffa [1957, p.90-91] described as husband and wife, have to decide whether they will go to a boxing match or to the ballet. They both prefer being together than being alone, but the husband prefers the boxing match while his wife prefers the ballet.

	Boxing	Ballet
Boxing	(2,1)	(0,0)
Ballet	(0,0)	(1,2)

Table 3.5: The Battle of the Sexes.

3.7.1. Fact. [Non-coordination in the Battle of the Sexe] Take the game of Figure 3.2 and assume that $X = \Pi_{i\in I}S_i$. Then for any pointed model $\mathbb{M}, w$ of that game if both agents have payoff-compatible intentions and are intention-rational at w then $\sigma(w) = Boxing - ballet$.

Proof. The only intention sets that are payoff-compatible are generated by $\{Boxing - Boxing\}$ for 1 and $\{Ballet - Ballet\}$ for 2. Agent 1 can thus only be intention-rational at states w where $\sigma(w)$ is either $Boxing - Boxing$ or $Boxing - Ballet$. Similarly, 2 can only be intention-rational at states w where $\sigma(w)$ is either $Ballet - Ballet$ or $Boxing - Ballet$. They can thus only be intention-rational with payoff-compatible intentions together at a state w if $\sigma(w) = Boxing - Ballet$. ■

To put it the other way around, this result shows that there cannot be intention-based coordination in the Battle of the Sexes if intentions are payoff-compatible. This is clearly due to the individualistic character of payoff-compatibility. It makes each agent intend to realise one or more of *his* best feasible outcomes, irrespective of what the others' intentions are. Intuitively, a more general account of intention-based coordination, one that does not rest on the specific structure of Hi-Lo games, will have to make the intentions of the agents more dependent on one another.

3.8 A general account of intention-based coordination

Mutual dependence of each others' intentions is the cornerstone of Bratman's sufficient conditions for "shared cooperative activity" [Bratman, 1999, p.105]. For my present purposes it is not necessary to go into his account in detail. Intention-based coordination is not necessarily a shared cooperative activity, as we shall see, but some of these requirements provide obvious anchors for coordination.

To start with, Bratman has emphasized the importance of *meshing sub-plans*. For him, "individual sub-plans concerning our [action] *mesh* just in case there is some way we could [do this action] that would not violate either of our sub-plans but would, rather, involve successful execution of those sub-plans."[Bratman, 1999, p.99] This is, in part, what goes wrong in the Battle of the Sexes. The agents have similar intentions, to achieve their most preferred feasible outcome, but their sub-plans—here their most precise intention—do not mesh. One excludes the other. Formally, the meshing sub-plans condition can be spelled out as follows[22]:

3.8.1. Definition. [Meshing sub-plans] The sub-plans of $i \in G \subseteq I$ *mesh* at a state w whenever $\bigcap_{i \in G} \downarrow\iota_i(w) \neq \emptyset$.

In other words, the sub-plans of agents in a group G mesh whenever they can be achieved together. Now, another important aspect of Bratman's account of shared cooperative activity is, indeed, that the agents involved have the intention to play their part.

3.8.2. Definition. [Intention agreement on A] The intentions of $i \in G \subseteq I$ *agree on* $A \subseteq X$ at w if $A \in \iota_i(w)$ iff $A \in \iota_j(w)$ for all $i, j \in G$.

Agents who agree on the intention to achieve A and whose sub-plans mesh already have "convergent" intentions. I shall write that these are "effectively" convergent whenever they do suffice to enforce an outcome in A.

3.8.3. Definition. [Effective intention convergence] The intentions of the agents $i \in G \subseteq I$ are *effectively convergent* at a state w in a given game model $\mathbb{M}$ if $\pi(\sigma(w)) \in A$ whenever all $i \in G \subseteq I$ agree on A and their sub-plans mesh at w.

Effective intention-convergence is just another way to say that agents who agree on achieving A by way of meshing sub-plans can do so. For an arbitrary strategic game $\mathbb{G}$, if C is the set of coordination points then agents whose intentions effectively converge on C are in principle able to coordinate. In other words, if the agents agree on the intention to coordinate, have meshing sub-plans to realise this intention and are effectively convergent, then we are sure they can coordinate. Before showing that precisely, it is important to see that intention-agreement, meshing sub-plans and effective convergence are all independent conditions.

3.8.4. Fact. [Independence (I)] There is a game $\mathbb{G}$ and a pointed model $\mathbb{M}, w$ of it where all agents have meshing sub-plans and their intentions agree on A, but they are not effectively convergent.

Proof. Take any game where each agent has two strategies and where $X = \Pi_{i \in I} S_i$. Take a model of it as in Figure 3.3, where $W = \Pi_{i \in I} S_i$. Fix $A = \{\sigma_2\}$ and $\downarrow\iota_i(\sigma_1) = A$ for all $i \in I$. Then at σ_1 the intentions of all agents agree on A, their sub-plans mesh but they are not effectively convergent. ■

3.8.5. Fact. [Independence (II)] There is a game $\mathbb{G}$ and a pointed model $\mathbb{M}, w$ of it where all the intentions are effectively convergent, the agents have meshing sub-plans but they do not agree on A.

[22] The idea of meshing sub-plans is of course more fit for extensive games, where one can make explicit the various plans of actions that each agent intends. The one I propose here is a transposition of this idea to the simpler structure of strategic games.

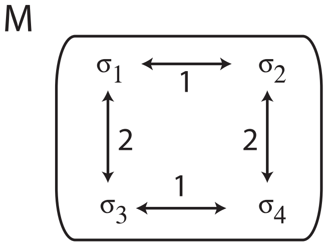

Figure 3.3: The model for the independence proofs.

Proof. Take the same set of states as in the previous proof. For one i, fix $\downarrow\iota_i(\sigma_1) = W$ and $\downarrow\iota_j(\sigma_1) = A$ for the other. We get that at σ_1 all agents have meshing sub-plans but they do not agree on A. The intentions then trivially satisfy effective convergence. ■

3.8.6. Fact. [Independence (III)] There is a game $\mathbb{G}$ and a pointed model $\mathbb{M}, w$ of it where all the intentions are effectively convergent, they agree on A but their sub-plan do not mesh.

Proof. Take the same game and set of states as in the previous proof, except that now take A to be $\{\sigma_1, \sigma_4\}$. For one i, fix $\downarrow\iota_i(\sigma_1) = \{\sigma_4\}$ and fix $\downarrow\iota_j(\sigma_1) = \{\sigma_1\}$ for the other. Then the intentions of all agents agree on A but their sub-plans do not mesh. Again, this means that they trivially satisfy effective convergence. ■

Taken together, meshing sub-plans, intention agreement and effective convergence are indeed sufficient for coordination.

3.8.7. Fact. [Intention-based coordination - the general case] Let $\mathbb{G}$ be a game let $C \subset \Pi_{i \in I} S_i$ be the non-empty set of coordination profiles.

- For any epistemic model $\mathbb{M}$ for $\mathbb{G}$ and any $w \in W$, if at w all agents' intentions agree on $\pi(C) = \{x : \exists \sigma \in C \text{ such that } \pi(\sigma) = x\}$, have meshing sub-plans and are effectively convergent, then $\sigma(w) \in C$.
- If $\sigma \notin C$ then we can construct a model $\mathbb{M}$ and a state w such that $\sigma(w) = \sigma$ and the above condition fails at w.

Proof. The first part is just unpacking the definitions. The models needed for the second part are easily adapted from the proofs of Facts 3.8.4, 3.8.5 and 3.8.6. ■

Fact 3.8.7 encompasses the three that we saw in Section 3.5. An easy check reveals that in each case the agents agree on the outcome of the Pareto-optimal profile, have meshing sub-plans and are effectively convergent. But, as one may have noticed, this does not ensure that the agents will enact their intentions. Meshing sub-plans, intention-agreement and effective convergence are independent of intention-rationality.

3.8.8. Fact. [Independence (IV)] There is a game $\mathbb{G}$ and a pointed model $\mathbb{M}, w$ where the intentions of all agents agree on A, are effectively convergent and are sub-plans meshing but some agents are intention irrational.

Proof. Take the same game and set of states as in the independence proofs above. Fix A as $\{\sigma_1, \sigma_4\}$ and set the intentions of all $i \in I$ to $\downarrow\iota_i(\sigma_1) = \{\sigma_4\}$. We get that the intentions of all agents agree on A, are effectively convergent and are sub-plans meshing but none of the agents is intention-rational. ■

This general account of coordination does not require the agent to know anything about the others' intentions. In fact, they do not even have to know that the three conditions hold or that they are at a coordination point. Coordination can thus occur in strategic game unbeknown to the agents. We already knew that from Fact 3.5.1, the first account of coordination in Hi-Lo games. This general, non-epistemic account of coordination shows that this can be the case for any game.

Just as I did in the case of Hi-Lo games, one can use Fact 3.8.7 as a starting point and strengthen it with various epistemic conditions. For example, mutual knowledge of intention agreement, meshing sub-plans and effective convergence are clearly enough to ensure coordination. Along these lines, it is interesting to see how "weak" are the sufficient conditions for coordination stated in Fact 3.8.7, in comparison with the conditions for shared cooperative activity that [Bratman, 1999, p.105] proposes. Among other things, he requires *common* knowledge of various conditions on intentions. This is a much stronger requirement than any epistemic conditions that we have encountered so far[23]. This does not mean that Bratman's condition are too strong, but rather that most cases of successful intention-based coordination in strategic games are not "shared cooperative activity" in his sense. In other words, intentions are indeed "all-purpose means" [Bratman, 2006a, p.275] for coordination. They can foster coordination not only in full-blown, cooperative, shared agency, but also in a very wide array of contexts.

[23] Once again it is not necessary to go into details of the definition of common knowledge. The interested reader can consult Fagin et al. [1995] and Aumann [1999] and van Ditmarsch et al. [2007] for details.

3.9 Conclusion

In this chapter we have seen that the theory of intention can legitimately claim a place among the theories that account for coordination in games. Using epistemic models, we have seen that intentions and mutual knowledge of intentions can foster coordination in the benchmark case of Hi-Lo games. Intention-based coordination, however, is not constrained to this particular class of strategic interaction. As we saw in the last section, one can easily spell out general conditions under which intentions anchor coordination in strategic games.

It should be observed that this last account of coordination *can* diverge from "standard" game-theoretical solutions. In Fact 3.8.7 I defined coordination profiles abstractly, without worrying whether these could be game-theoretically rationalized. But one could clearly impose further restrictions on the coordination points, as I did with payoff-compatibility, in order to make them fit other rationality requirements.

Along these lines, it is worth recalling Sen's [2005] famous claim that intentions (or more generally commitments) are of interest mainly when they do *not* coincide with standard rationality. The results of section 3.5 show that intentions can be of interest *even when* they coincide with classical notions of rationality. The general account of Section 3.7, however, allows for cases that are in line with Sen's view, that is cases where intentions do not coincide with standard rationality. I have not looked at such cases in detail here, but they surely deserve more scrutiny.

There is also much more to be explored on the connection between intention-based coordination and other accounts in the game-theoretic literature. Here I have only examined one such, the Stackelberg Heuristic, because of its obvious connection with payoff-compatibility of intentions, but the intention-based account should be compared with other proposals such as Bacharach's [2006] group-oriented reasoning or the more classically correlated beliefs of Aumann [1987]. The latter is especially interesting, in view of the general account of coordination of Section 3.7. It would be illuminating to see whether knowledge of intention agreement, for example, could serve as a basis for correlated belief systems in strategic games.

This, of course, would lead towards a more *belief-based* analysis, which would surely deepen our understanding of intention-based reasoning in games. But the results we have so far already show that intentions, even in a knowledge-based perspective, *can* foster coordination in games. This is in itself valuable both from the point of view of the theory of intentions and from a game-theoretical perspective. In the next chapter I turn to another important role of intentions in practical reasoning, the reasoning-centered commitment.

3.10 Appendix - Solution concepts

As mentioned in the Section 1.1.2 of the introduction, in game theory there are many ways to understand instrumental rationality. Different solution concepts encapsulate these different understandings. Here I present the formal definitions for the two that I use in the body of the text. For further explanations of them see Myerson [1991].

3.10.1. Definition. [Dominated strategy] In a given strategic game $\mathbb{G}$, a strategy $s_i \in S_i$ is *strictly dominated* by $s'_i \in S_i$ iff $(s'_i, \sigma_{j\neq i}) \succ_i (s_i, \sigma_{j\neq i})$ for all $\sigma_{j\neq i}$. It is *weakly dominated* by s'_i iff $(s'_i, \sigma_{j\neq i}) \succeq_i (s_i, \sigma_{j\neq i})$ for all $\sigma_{j\neq i}$ but there is one $\sigma'_{j\neq i}$ such that $(s'_i, \sigma_{j\neq i}) \succ_i (s_i, \sigma_{j\neq i})$.

3.10.2. Definition. [Removal of strictly dominated strategies] The game $SD(\mathbb{G})$ which results after *elimination of strictly dominated strategies* from $\mathbb{G}$ is defined as follows[24]:

- $SD(S_i) = \{s_i : s_i \text{ is not strictly dominated by some other } s'_i \in S_i\}$.
- $X^{SD} = \{x \in X : x = \pi(s_i) \text{ for some } s_i \in SD(S_i)\}$.
- π^{SD} is the restriction of π to $SD(S_i)$.
- $\succeq_i^{SD}$ is the restriction of $\succeq_i$ to X^{SD}.

The game $SD^\omega(\mathbb{G})$ which results after *iterated* removal from $\mathbb{G}$ of strictly dominated strategies is inductively defined as follows: $SD^\omega(\mathbb{G}) = \bigcap_{n<\omega} SD^n(\mathbb{G})$ where $SD^0(\mathbb{G}) = \mathbb{G}$ and $SD^{n+1}(\mathbb{G}) = SD^n(\mathbb{G})$.

The "rational" strategies according to this solution concept are those which survive iterated removal, i.e. those $s_i \in SD^\omega(S_i)$. For an in-depth investigation of the formal properties of this solution concept, and many others, see Apt [2007].

3.10.3. Definition. [Pure Nash equilibrium] A strategy profile σ is a *pure Nash equilibrium* iff $(\sigma(i), \sigma_{j\neq i}) \succeq_i (s'_i, \sigma_{j\neq i})$ for all $i \in I$ and $s'_i \in S_i$.

A Nash equilibrium is a profile where all players play their *best response*, see Section 3.6, given the strategy choices of the others. Osborne and Rubinstein [1994] offer a formal review of the properties of Nash equilibria. See also Myerson [1991].

[24] Removal of weakly dominated strategies is more tricky to define. See Myerson [1991, p.90].

Chapter 4

Intentions and transformations of strategic games

We saw in the previous chapter how intentions, by committing agents to action, can anchor coordination. Planning agents can better anticipate their own choices as well as those of others. From this point of view intentions play a passive but nevertheless crucial role in practical reasoning. They are facts that planning agents use as the basis for their deliberations. But the planning theory has it that intentions also play a more active part in deliberation. They *shape* decision problems. To quote Bratman [1987, p.33, emphasis in original]:

> My prior intentions and plans [...] pose problems for deliberation, thereby establishing standards for *relevance* for options considered in deliberation. And they constrain solutions to these problems, providing a *filter of admissibility* for options. They narrow the scope of the deliberation to a limited set of options. And they help answer a question that tends to remain unasked within traditional decision theory, namely: where do decision problems come from?

Intentions influence the way agents look at decision problems by imposing a "standard for relevance" and a "filter of admissibility" on options. The latter comes from the consistency requirements I presented in the Introduction (Section 1.2). It should be possible to choose an option without contradicting what the agent already intends. Intentions rule out such contradictory options, hence the idea of a filter of admissibility. The standard of relevance stems, on the other hand, from the norm of means-end coherence. Intentions spur practical reasoning towards deliberation on means. They "pose problems for *further* deliberations," where "relevant" options are those that foster intended goals.

This influence on the very content of decision problems, the *reasoning-centered commitment* of intentions, is the subject of this chapter. Section 4.1 is about ruling out options that are inadmissible in view of previously adopted intentions, and Section 4.2 is about means-end coherence and its standard for relevance. In Section 4.3 I connect these two ways to transform decision problems.

I am mainly interested in reasoning-centered commitment for agents in situations of strategic interaction. The benefits of the reasoning-centered commitment for (resource-bounded) *individuals* have been made explicit by Bratman et al. [1991], Pollack [1992] and Horty and Pollack [2001]. Briefly, intentions simplify and focus deliberation, avoiding strenuous pondering. Very little attention has been paid, however, to reasoning-centered commitment in the context of games. As we shall see, interactive contexts unveil a new dimension to this function of intentions. It is no longer a matter of "simply" fitting one's options with what one intends. The presence of other planning agents forces one to take *their* intentions into account, even in games where there is at first sight no incentive to do so. Bringing in reasoning-centered commitment at the level of games thus poses new problems about rational interaction. To paraphrase Bratman [1987, p.33], these problems are generally left unaddressed in both game theory and the theory of intentions.

In what follows I use the same definitions of *strategic games* and *intention sets* as in the previous chapter[1]. I now examine how intentions can transform strategic games and their various components, namely strategies, outcomes and preferences. I do not, however, look at how these transformations affect what agents might know about each other in games. That is, I do not look at transformations of *epistemic models* for games, which will have to wait until Chapter 5, where logical methods are introduced that facilitate the analysis.

4.1 Ruling out options

As just mentioned, by "ruling out options" I mean imposing the filter of admissibility which stems from the consistency requirements on intentions. In single-agent decision problems, to rule out options is just to remove inadmissible strategies, viz., those that do not yield any intended outcome. One can define admissibility in many ways in multi-agent contexts, though. An agent i may or may not take into account the fact that he is interacting with other planning agents while filtering his set of options. He may also be more or less willing to run the risk of not reaching an intended outcome. Let me therefore start with a generic definition of filtering, to which I attach different notions of admissibility and compare their respective behaviour.

4.1.1. Definition. [Cleaned strategy set] The *cleaned version* $cl(S_i)$ of a strategy set S_i is defined as:

$$cl(S_i) = \{s_i \mid s_i \text{ is admissible for deliberation for } i\}$$

[1] Strategic games and intentions sets are defined on page 36 and 33, respectively.

4.1.2. Definition. [Cleaning of strategic games] The *cleaned* version of a game $\mathbb{G}$ with intention profile ι is the tuple $cl(\mathbb{G}) = \langle I, X^{cl}, \{cl(S_i), \succeq_i^{cl}\}_{i \in I}, \pi^{cl} \rangle$ such that:

- $X^{cl} = \pi(\Pi_{i \in I} cl(S_i)) = \{x \mid x = \pi(\sigma) \text{ for some } \sigma \in \Pi_{i \in I} cl(S_i)\}$.
- $\succeq_i^{cl}$ is the restriction of $\succeq_i$ to X^{cl}.
- π^{cl} is π with the restricted domain $\Pi_{i \in I} cl(S_i)$.

The cleaned version ι_i^{cl} of the intention set ι_i for agent i is the filter generated by $\downarrow\iota_i \cap X^{cl}$.

The cleaned version of a strategic game is thus the game that results from looking only at admissible strategy profiles, whatever "admissible" means. The outcome set is reduced accordingly, i.e. to the outcomes that can result from admissible profiles. The preference relations are, in turn, reduced to the cleaned outcome set.

In this definition, cleaning also modifies the intentions of the agents. The cleaned ι_i^{cl} is the restriction of ι_i to the outcomes that survive the cleaning. The idea here is that the agents adapt their intentions to the new decision problem they face after cleaning, in a very down-to-earth manner. They simply give up on achieving the outcomes that are no longer achievable in the reduced game. The consequences of such a way of adapting one's intentions depend heavily on how admissibility is defined.

Intention rationality, introduced in the previous chapter for epistemic models, provides a natural starting point to think about this notion[2]. Recall that an agent is intention-rational when he plays a strategy s_i which could yield at least some outcome in his intention set ι_i. Conversely, the choice of s_i is intention-irrational if it makes all intended outcomes impossible. This idea is easily conveyed from strategy choices to admissibility of strategies.

4.1.3. Definition. [Individualistic admissibility] A strategy s_i of agent i is *individualistically admissible* with respect to his intention set ι_i when $\pi(s_i) \cap \downarrow\iota_i \neq \emptyset$.

In other words, a strategy is individualistically admissible for deliberation when choosing it could yield at least one outcome $x \in \downarrow\iota_i$. I call this notion "individualistic" because it is related only to the agent's own intentions[3].

[2] I show in Chapter 5 that there is in fact a formal connection between intention-rationality and cleaning with "altruistic admissibility", which I introduce shortly. The second is so to speak the dynamic expression of the first and, vice-versa, the first is the static counterpart of the second.

[3] I should stress that the claim here is not that cleaning with individualistic admissibility exactly captures what Bratman meant by "filter of admissibility" in the quote at the beginning of the chapter. The operation I study here, as well as those that follow in this chapter, are at best intended to represent various dimensions of the reasoning-centered commitment.

Even in the single agent case, cleaning with individualistic admissibility has interesting consequences when it is coupled with the intention adaptation mechanism from Definition 4.1.2. Observe that, in general, not all outcomes need to be *realizable* by a strategy profile in a given strategic game. There can be outcomes $x \in X$ for which there is no profile σ such that $\pi(\sigma) = x$. But the intention sets are defined only with respect to the set of outcomes X. Nothing thus precludes an agent i from having some unrealizable outcomes in his most precise intention $\downarrow\iota_i$[4]. In that case, cleaning with individualistic admissibility brings the agent in tune with reality, so to speak. It eliminates the unrealizable outcomes from his intention set.

It can happen, however, that an agent i has *only* unrealizable outcomes in $\downarrow\iota_i$. In that case cleaning leaves *no* admissible strategy to choose. Cleaning thus means that agents can be confronted with the unrealistic character of their own intentions. Intuitively, in these cases the agents would have to *revise* their intentions. I do not, however, venture into the realm of intention revision here. The issue returns in a different guise in Chapter 5, where I use a minimal revision policy. For the present I will rather focus on "*ex ante*" conditions on intended outcomes.

For single agent decision problems, these conditions are easy to pinpoint. Cleaning makes a decision problem empty if and only if the agent does not intend any realizable outcomes. But in interactive situations, agents who clean individualistically can, by ruling out some of their own strategies, cause the outcomes that others intend to be unrealizable. Consider, for example, Table 4.1, with the numbers in the cells representing which outcomes are in $\downarrow\iota_i$ for the corresponding agent. Here the only admissible strategy for agent 2 is t_1, because he intends the outcome of (s_2, t_1). But observe that this very outcome gets ruled out by 1's cleaning. From his point of view, s_2 is not an admissible strategy, because he only intends to achieve the outcome of (s_1, t_1). After one round of cleaning, there is thus no way in which agent 2 can achieve his original intention. Following the intention adaptation rule from Definition 4.1.1, he thus ends up with an empty $\downarrow\iota_i^{cl}$, which means that his intention set is not internally consistent.

$\mathbb{G}$	t_1	t_2
s_1	1	
s_2	2	

$cl(\mathbb{G})$	t_1
s_1	1

Table 4.1: A game which an empty cleaning.

[4] I ignored the possibility of unrealizable outcomes in the previous chapter because most of the intention sets contained, by definition, realizable outcomes. The reference to feasible outcome in the definition of payoff-compatible intentions, for example, ensured realizability. Similarly, effective intention convergence had realizability built in.

Cleaning, in such cases, makes the game[5] *empty* because, after sufficient rounds, there is an agent i such that $cl(S_i) = \emptyset$. In the example above, the game becomes empty at the second step of cleaning, because after the first step agent 2 ends up with internally inconsistent intentions. This holds in general.

4.1.4. Fact. [Empty cleaning and internally inconsistent intention sets] For all strategic game $\mathbb{G}$ and intention profile ι the following are equivalent with individualistic cleaning.

- $cl(\mathbb{G})$ is empty.
- There is an agent i such that for all $x \in \downarrow\iota_i$, x is not realizable in $\mathbb{G}$.

Proof. Let X^* be the set of realizable outcomes in $\mathbb{G}$. Observe that for any agents i, $\bigcup_{s_i \in S_i} \pi(s_i) = X^*$. The equivalence follows directly from this fact: $cl(\mathbb{G})$ is empty iff there is a i such that $cl(S_i) = \emptyset$, which by definition happens iff for all $s_i \in S_i$ we have $\pi(s_i) \cap \iota_i = \emptyset$. This, by the above observation, happens iff no $x \in \iota_i$ is in X^*. ■

This means that, as in the example above, if all agents have intentions that are realizable in the original game then cleaning needs at least two steps to reach an empty point. In fact, in most cases cleaning needs more than one step to reach a sub-game which does not reduce any further. Of course, not all strategic games become empty after cleaning. To take another example, consider Table 4.2. In the original game, the leftmost table, only 2 has an individualistically inadmissible strategy, namely t_2. But by ruling out this strategy he also excludes the only outcome that makes s_2 admissible for agent 1. Adapting his intention set accordingly, 1 no longer considers s_2 admissible after the first round of cleaning (the centre matrix). One more cleaning round thus results in the rightmost table, where all strategies are admissible.

$\mathbb{G}$	t_1	t_2
s_1	1, 2	
s_2		1

$cl(\mathbb{G})$	t_1
s_1	1, 2
s_2	

$cl(cl(\mathbb{G}))$	t_1
s_1	1, 2

Table 4.2: A game where individualistic cleaning stops after two steps.

The last examples feature two crucial aspects of cleaning with individualistic admissibility, namely the notion of a *fixed-point* and the possibility of empty

[5] Here I slightly abuse the terminology, because by definition for a structure to be a strategic game all S_i should be non-empty.

cleaning. To study these phenomena in full generality, let me introduce some more notation[6].

4.1.5. Definition. [Iterated cleaning] Given a strategic game $\mathbb{G}$, let $cl^k(\mathbb{G}) = \langle I, X^{cl^k}, \{cl^k(S_i), \succeq_i^{cl^k}\}_{i\in I}, \pi^{cl^k}\rangle$ be the strategic game that results after k iterations of the cleaning of $\mathbb{G}$. That is, $cl^1(\mathbb{G}) = cl(\mathbb{G})$ and $cl^{k+1}(\mathbb{G}) = cl(cl^k(\mathbb{G}))$. The smallest[7] cleaning *fixed-point* $cl^{\#}(\mathbb{G})$ of $\mathbb{G}$ is defined as $cl^k(\mathbb{G})$ for the smallest k such that $cl^k(\mathbb{G}) = cl^{k+1}(\mathbb{G})$.

Obviously, every game has a unique cleaning fixed point with individualistic cleaning. This follows directly from the fact that, first, I work with finite games, second, that the cleaned version of a strategic game is always one of its sub-games, and finally, that each game has a unique cleaned version[8]. More interestingly, games have *non-empty* cleaning fixed points whenever the agents' intentions are sufficiently entangled.

4.1.6. Definition. [Cleaning core] The *cleaning core* of a strategic game $\mathbb{G}$ is the set of strategy profile S^* inductively defined as follows, with $\pi^{S^n}(s_i) = \pi(s_i) \cap \{\pi(\sigma') : \sigma' \in S^n\}$.

- $S^0 = \Pi_{i\in I} S_i$.
- $S^{n+1} = S^n - \{\sigma :$ there is an i such that $\pi^{S^n}(\sigma(i)) \cap \downarrow\iota_i = \emptyset\}$.
- $S^* = \bigcap_{n\leq\omega} S^n$.

From the perspective of each agent, the cleaning core is a set of strategies $S_i^* \subseteq S_i$ that are very tightly connected to what the other agents intend. For each strategy s_i and profile σ in the cleaning core such that $\sigma(i) = s_i$, there is at least one agent j for whom strategy $\sigma(j)$ is admissible, *by looking only at what can result from the profiles in the core.* Furthermore, it follows from this definition that there has to be at least one of these σ that yields an outcome that i himself intends. Unsurprisingly, S^* is not empty for a given strategic game precisely when this game has a non-empty cleaning fixed point.

[6] The process of iteration that I define here is quite similar in structure to the one of iterated elimination of dominated strategies (see Chapter 3, Appendix and van Benthem [2003]). I do not investigate the connection between the two processes. For example it might be that iterated elimination of dominated strategies could be reproduced using the cleaning operation and a notion of admissibility which is sensitive to preferences. I do, however, look more carefully at how the two operations interact in Chapter 5, Section 5.3.3.

[7] In most of what follows I will ignore the "smallest" and only write about the fixed point.

[8] This observation, as well as many others in this chapter, are direct consequences of the underlying mathematical properties of cleaning and what I later call clustering. As pointed out by Apt [2007], monotonicity of these two operations ensures, for instance, the existence of a fixed point. See Fact 5.3.16 in Chapter 5 for the definition of monotonicity. My focus here is rather on the existence of *non-empty* fixed points.

4.1.7. Fact. [Non-empty fixed points] For any strategic game $\mathbb{G}$ and intention profile ι, the following are equivalent.

1. $S^* \neq \emptyset$.
2. $cl^{\#}(\mathbb{G})$ is not empty.

Proof. By Definition 4.1.6, (1) is the same as saying that we can find a $\sigma \in S^*$ such that for all i, $\pi^{S^*}(\sigma(i)) \cap \downarrow\iota_i \neq \emptyset$. I show by induction that $\pi(S^k) = X^{cl^k}$, for all k. This is enough to show the equivalence, for then we know that $X^{cl^{\#}} \cap \downarrow \iota_i \neq \emptyset$, which we know is the same as $cl^{\#}(\mathbb{G})$ being non-empty, from Fact 4.1.4. The basic case of the induction, $k = 0$, is trivial. For the induction step, assume the claim is proved for k. We have that $x \in \pi(S^{k+1})$ iff there is a $\sigma \in S^{k+1}$ such that $\pi(\sigma) = x$. This in turns happens iff $\pi^{S^k}(\sigma(i)) \cap \iota_i \neq \emptyset$, for all i. But by our inductive hypothesis this is just to say that $\pi(\sigma(i)) \cap X^{cl^k} \cap \iota_i \neq \emptyset$, which is just the definition of x being in X^{x+1}. ■

Fact 4.1.7 tells us that the individualistic character of admissibility must be compensated by an interlocking web of intentions and strategies if cleaning is not to make the game empty. Indeed, each strategy in the cleaning core is tightly connected with what *all* agents intend. Or, conversely, intentions which yield a non-empty cleaning core closely fit the admissible strategies of all agents. By intending outcomes that are realizable in the cleaning core, an agent somehow acknowledges that he interacts with other planning agents who, like him, clean inadmissible options from their strategy set[9].This can be appreciated even better by considering an alternative form of admissibility, which I call *altruistic.*

4.1.8. Definition. [Altruistic admissibility] A strategy s_i of agent i is *altruistically admissible* with respect to his intention set ι_i when there is a $j \in I$ such that $\pi(s_i) \cap \downarrow\iota_j \neq \emptyset$.

Following this alternative criterion, a strategy of agent i is admissible whenever it can yield an outcome that some agent, *not necessarily* i, intends. Agents here clean their strategy sets with an explicit concern for their co-players. This turns out to be enough to prevent empty cleanings, because it can no longer happen that agents make some outcomes intended by others unrealizable. After one round of cleaning all strategies are altruistically admissible.

4.1.9. Fact. [Fixed point for altruistic admissibility] For $\mathbb{G}$ an arbitrary strategic game, $cl^{\#}(\mathbb{G}) = cl(\mathbb{G})$ for cleaning with altruistic admissibility.

[9] In the absence of an epistemic analysis of non-empty cleaning, this claim is bound to remain vague, hence my use of "somehow". Intuitively, however, it seems quite clear that for agents to acknowledge or to take into account the intentions of other they would have to have some *information*, i.e. knowledge or beliefs, about them. This is precisely what an epistemic analysis could provide.

Proof. I show that $cl(cl(\mathbb{G})) = cl(\mathbb{G})$. Given the definition of the cleaning operation, it is enough to show that $cl(cl(S_i)) = cl(S_i)$ for all i. It should be clear that $cl(cl(S_i)) \subseteq cl(S_i)$. It remains to show the converse. So assume that $s_i \in cl(S_i)$. Since cleaning is done with altruistic admissibility, this means that there is a σ such that $\sigma(i) = s_i$ and a $j \in I$ such that $\pi(\sigma) \in \downarrow\iota_j$. But then $\sigma(i') \in cl(S_{i'})$ for all $i' \in I$, and so $\sigma \in \Pi_{i \in I} cl(S_i)$. This means that $\pi(\sigma) \in X^{cl}$, which in turns implies that $\pi^{cl}(\sigma) \in \downarrow\iota_j^{cl}$. We thus know that there is a $\sigma \in \Pi_{i \in I} cl(S_i)$ such that $\sigma(i) = s_i$ and a j such that $\pi^{cl}(\sigma) \in \downarrow\iota_j^{cl}$, which means that $s_i \in cl(cl(S_i))$. ■

4.1.10. Fact. [Non-empty cleaning with altruistic admissibility] For any strategic game $\mathbb{G}$ and intention profile ι, the following are equivalent for cleaning with altruistic admissibility.

- For all i, there is a realizable $x \in\downarrow\iota_i$.
- $cl^{\#}(\mathbb{G})$ is not empty.

Proof. There is a realizable $x \in \iota_i$ for all i iff for all i there is a σ such that $\pi(\sigma) \in \downarrow\iota_i$. But this is this same as to say that for all j there is a strategy s_j such that $\sigma(j) = s_j$ and an i such that $\pi(\sigma) \in \downarrow\iota_i$ which, by Facts 4.1.7 and 4.1.9, means that $cl^{\#}(\mathbb{G})$ is not empty. ■

This shows even more clearly how crucial it is for agents to take the others' intentions into account when ruling out options in strategic games. If, on the one hand, agents rule out options without taking care of what the others intend, they run the risk of ending up with no strategy at all, unless their intentions are already attuned to those of their co-players. If, on the other hand, their intentions do not fit so well with those of others, then they should at least take heed of what the others intend when ruling out options.

As I mentioned at the beginning of this section, there are many other ways to define admissibility. Here I have looked at two variants in which the agents care to a different extent about the intentions of others. In both cases a strategy is admissible if in some scenario, maybe only *one*, it yields an intended outcome. But if a huge number of scenarios are compatible with a single choice of strategy, the agent might be more careful in assessing admissibility. He might, for instance, only consider admissible strategies which yield intended outcomes in a majority of cases. I do not investigate here what cleaning would look like with such a criterion. Rather, I now move to the second aspect of reasoning-centered commitment, namely the standard of relevance that stems from intentions. In Chapter 5 I return to (altruistic) cleaning, but this time to study how it transforms the information that agents have about each other in strategic games.

4.2 Grouping redundant options

Intentions impose a standard for relevance of options because planning agents are under rational pressure to form intentions about means to achieve what they intend. Some options might not be considered relevant simply because they are no means to achieve one's intentions. This is what I have studied in the last section. But even among the admissible options, there might be differences that are not relevant with respect to achieving one's end. In other words, some options might just be *redundant* in terms of the agent's intentions. Look for example at the game in Table 4.3, again with the numbers 1 and 2 in the cells referring to the outcomes that are in $\downarrow\iota_1$ and $\downarrow\iota_2$. Observe that agent 1 gets an intended outcome in the exact same circumstances by choosing s_1 or s_2. In both cases, he obtains an intended outcome if 2 chooses t_1 or t_3, but not if 2 chooses t_2. If 1 looks at his options as ways to satisfy his intentions, there is no significant difference between s_1 and s_2. In view of choosing a means, strategies s_1 and s_2 are *redundant*. He could just as well treat them as *one* way to achieve what he intends, thus discarding irrelevant details. The following notion of redundancy embodies such a standard of relevance.

	t_1	t_2	t_3
s_1	1, 2	2	1
s_2	1	2	1
s_3		1	2

Table 4.3: A game with two means-redundant strategies for agent 1.

4.2.1. Definition. [Means-redundancy] Two strategies s_1 and s_2 in S_i are *means-redundant*, noted $s_1 \approx s_2$, whenever $\pi(s_1, \sigma_{j\neq i}) \in \downarrow\iota_i$ iff $\pi(s_2, \sigma_{j\neq i}) \in \downarrow\iota_i$ for all combinations of actions of other agents $\sigma_{j\neq i} \in \Pi_{j\neq i} S_j$.

Means-redundant options are thus options which yield intended outcomes in exactly the same circumstances. Options that are not means-redundant, on the other hand, are genuinely distinct means to achieve what one intends. They are different ways to meet the means-end coherence requirement. This idea is naturally captured by the fact that $\approx$ is an equivalence relation. It partitions the set of strategies S_i into subsets $[s_i]^{\mathbb{G}}_{\approx} = \{s'_i \in S_i | s'_i \approx s_i\}$, each of which represents a distinct means for agent i to achieve what he intends. The clustering of redundant options thus gives a "means-oriented" perspective on decision problems.

But to make a decision from this means-oriented perspective, the agents need to *evaluate* these means, i.e., to assess which one they prefer, and form expectations about how the others will evaluate theirs. Here I use the underlying preference ordering on outcomes, by assuming that each agent "picks" according to *some* criterion one strategy per means $[s_i]^{\mathbb{G}}_{\approx}$, and collects these picked strategies to form his new strategy set. The idea here is that agents might still acknowledge differences between two strategies, even though these differences are not relevant from a means-oriented point of view. But by picking *inside* clusters of strategies, planning agents give priority to decision on means. They first sort out their options with respect to what they intend. Only then, among the options that are equivalent means, do they invoke other discriminating criteria. In other words, the focus on means lexicographically precedes any other decision rules.

A lot of different criteria might drive this picking. Rational expectations and preferences are obvious candidates, but these are not the only ones[10]. To keep the analysis as general as possible I use the following abstract definition of *picking functions.*

4.2.2. Definition. [Picking function] Given a strategic game $\mathbb{G}$, a function $\theta_i : \mathcal{P}(S_i) \to S_i$ such that $\theta_i(S) \in S$ for all $S \subseteq S_i$ is called i's *picking function.* A *profile* of picking functions Θ is a combination of such θ_i, one for each agent $i \in I$.

These functions thus return, for each set of strategies—and in particular each equivalence class $[s_i]_{\approx}$—the strategy that the agents picks in that set. I define them over the whole power set of strategies, instead of over the sets $[s_i]^{\mathbb{G}}_{\approx}$ because it makes the technical details much simpler in what follows. As mentioned, one can constrain these functions in various ways, to encode different picking criteria. These lead to different *prunings* of strategy sets.

4.2.3. Definition. [Pruned Strategy set] The *pruned version* $pr(S_i)$ of a strategy set S_i, with respect to an intention set ι_i and a picking function θ_i is defined as:

$$pr(S_i) = \{\theta([s_i]^{\mathbb{G}}_{\approx}) : s_i \in S_i\}$$

For cleaning, admissibility provided the criterion for transforming each agent's strategy set, and from there I defined the corresponding transformation of strategic games. The situation is entirely similar here, except that the transformation of the strategy set proceeds in two steps. First, the agents group their options into different means to achieve what they intend. They then pick one option per means, according to whatever criterion θ_i encodes. The strategic games which result from this two-step transformation are defined in exactly the same fashion as those which result from cleaning.

4.2.4. Definition. [Pruning of strategic games] The *pruned version* of a strategic game $\mathbb{G}$, from the perspective of an intention profile ι and of a profile of picking function Θ is the tuple $pr(\mathbb{G}) = \langle I, X^{pr}, \{pr(S_i), \succeq_i^{pr}\}_{i\in I}, \pi^{pr}\rangle$ such that:

[10] To push the investigation further in that direction one could look at the work of Jehiel and Samet [2003] or in the literature on social choice and preference or judgment aggregation, for example in the classical works of Arrow [1970] and Sen [1970]. For further references on judgment aggregation see List [2007].

- $X^{pr} = \pi(\Pi_{i \in I} pr(S_i))$.
- $\succeq_i^{pr}$ is the restriction of $\succeq_i$ to X^{pr}.
- π^{pr} is π with the restricted domain $\Pi_{i \in I} pr(S_i)$.

The pruned version ι_i^{pr} of an intention set ι_i is the filter generated by $\downarrow\iota_i \cap X^{pr}$.

This definition once again features the idea that agents should adapt their intentions in the process of pruning. They abandon achieving the outcomes that are no longer realizable. In the case of cleaning this opened the possibility for agents to end up with internally inconsistent intentions, and as a consequence for strategic games to have empty cleaned fixed points. The situation is similar in the case of pruning. It can happen that agents end up with internally inconsistent intentions after a few rounds of pruning.

Consider for example the leftmost matrix in Table 4.4, and suppose that $\theta_1(\{s_1, s_2\}) = s_2$ and $\theta_1(\{s_2, s_3\}) = s_2$ for agent 1 and $\theta_2(\{t_1, t_2\}) = t_2$ and $\theta_2(\{t_2, t_3\}) = t_2$ for agent 2. The picking criterion of each agent removes all the intended outcomes of the other, leaving them with empty intention sets after one step of pruning. No more pruning can reduce the matrix on the right of Table 4.4. It is the *pruning fixed point* of $\mathbb{G}$.

$\mathbb{G}$	t_1	t_2	t_3
s_1	1, 2	2	
s_2	1		
s_3			

$pr(\mathbb{G})$	t_2	t_3
s_2		
s_3		

$pr(pr(\mathbb{G}))$	t_2
s_2	

Table 4.4: A game in which pruning removes all intended outcomes.

4.2.5. Definition. [Iterated pruning] Given a strategic game $\mathbb{G}$, let $pr^k(\mathbb{G})$ be the strategic game that results after k iterations of the pruning of $\mathbb{G}$. That is, $pr^0(\mathbb{G}) = \mathbb{G}$ and $pr^{k+1}(\mathbb{G}) = pr(pr^k(\mathbb{G}))$. The *pruning fixed point* $pr^{\#}(\mathbb{G})$ of $\mathbb{G}$ is defined as $pr^k(\mathbb{G})$ for the smallest k such that $pr^k(\mathbb{G}) = pr^{k+1}(\mathbb{G})$.

The pruning fixed point in the above table has two interesting features which generalize to arbitrary strategic games. Observe first that even though both agents have internally inconsistent intentions in $pr^{\#}(\mathbb{G})$, this does not lead to an empty game. Pruning, in fact, *never* make strategic games empty.

4.2.6. Fact. [Non-empty pruning] For all strategic game $\mathbb{G}$ and agent $i \in I$, $pr^{\#}(S_i) \neq \emptyset$.

Proof. This is shown by induction on $pr^k(\mathbb{G})$. The basic case is trivial. For the induction step, observe that the picking function θ_i is defined for the whole power set of S_i. This means, given the inductive hypothesis, that $\theta_i([s_i]_{\approx}^{pr^k(\mathbb{G})})$ is well-defined and in $[s_i]^{pr^k(\mathbb{G})}$ for any $s_i \in pr^k(S_i)$, which is enough to show that $pr^{k+1}(S_i)$ is also not empty. ■

For any game $\mathbb{G}$, as in the example above, it is also worth noting that there is no means-redundancy at the pruning fixed point $pr^{\#}(\mathbb{G})$. All options are genuinely different from the means-oriented perspective. This is indeed what being a fixed point means. For all agents i, all sets $[s_i]_{\approx}^{pr^{\#}(\mathbb{G})}$ are singletons and so $\theta_i([s_i]_{\approx}^{pr^{\#}(\mathbb{G})}) = s_i$. There is no way to reduce the strategy sets further.

For cleaning with individualistic admissibility, the existence of a non-empty fixed point rests on a tight connection between the agents' intentions. The situation is similar here. The existence of pruning fixed points where all agents have consistent intentions depends on whether they intend "safe" outcomes.

4.2.7. Definition. [Safety for pruning] Given a strategic game $\mathbb{G}$, an intention profile ι and a profile of picking functions Θ, the outcome $x = \pi(\sigma)$ is:

- *Safe for pruning at stage 1* iff for all agents i, $\theta_i([\sigma(i)]) = \sigma(i)$.
- *Safe for pruning at stage* $n + 1$ whenever it is safe for pruning at stage n and for all agents i, $\theta_i([\sigma(i)]^{pk^n(\mathbb{G})}) = \sigma(i)$.
- *Safe for pruning* when it is safe for pruning at all stages n.

The picking functions θ_i are the cornerstones of this inductive definition. Safe outcomes are those which the function retains, whatever happens in the process of pruning. It should thus not come as a surprise that intending safe outcomes is necessary and sufficient for an agent to keep his intention set consistent in the process of pruning.

4.2.8. Fact. [Intention-consistency at $pr^{\#}(\mathbb{G})$] *For any strategic game $\mathbb{G}$, intention profile ι and profile of picking function Θ, the following are equivalent for all $i \in I$.*

1. $\downarrow\iota_i^{pr^{\#}} \neq \emptyset$
2. There is a $\pi(\sigma) \in\downarrow\iota_i$ safe for pruning in $\mathbb{G}$.

Proof. From (1) to (2). Take any $x \in\downarrow\iota_i^{pr^{\#}}$. By definition we know that there is a $\sigma \in \Pi_{i\in I} pr^{\#}(S_i)$ such that $\pi(\sigma) = x$. But this happens iff $\sigma \in \Pi_{i\in I} pr^k(S_i)$ for all k, and so that $\theta_i([\sigma(i)]_{\approx}^{pr^k(\mathbb{G})}) = \sigma(i)$ also for all k, which in turns means that x is safe for pruning in $\mathbb{G}$. From (2) to (1), take any such $\pi(\sigma) \in \downarrow\iota_i$. I will show that $\pi(\sigma) \in X^{pr^k}$ for all k. The basic case is trivial, so assume that $\pi(\sigma) \in X^{pr^k}$. We know by definition that $\pi(\sigma)$ is safe for pruning at k, which gives automatically that $\pi(\sigma) \in X^{pr^{k+1}}$. ■

If pruning is not to lead the agents into internally inconsistent intentions, they are required to take the others' intentions *and* picking criteria into account[11]. Indeed, the notion of safe outcome for an agent i crucially involves both the intentions and the picking function of *all* agents in a given strategic game. A quick check reveals that in single-agent cases pruning never makes the intention set of the agent internally inconsistent, as long as the agent has realizable intentions. This shows, once again, that reasoning-centered commitment really gains an interactive character in situations of strategic interaction.

Another way to appreciate this fact is to compare the result of pruning a given game with *different* picking functions. Consider for example the game in Table 4.5. Assume that there is a one-to-one correspondence between profiles and outcomes, and that the agents have the following intentions : $\downarrow\iota_1 = \{(s_1, t_1), (s_2, t_1)\}$ and $\downarrow\iota_2 = \{(s_1, t_2), (s_2, t_2)\}$. Agent 1 has two ways to prune his options, because $s_1 \approx s_2$. Either $\theta_1([s_1]) = s_1$ or $\theta'_1([s_1]) = s_2$. Agent 2, on the other hand, has no means-redundant strategy. The games resulting from pruning with the picking functions of agent 1 are displayed in Table 4.6. Clearly, by picking according to θ'_1 agent 1 does not take 2's preferences into account. He can rationally expect 2 to choose t_2 in the game pruned with θ'_1, from which he becomes strictly worse off than in the game pruned by θ_1.

$\mathbb{G}$	t_1	t_2
s_1	(1,2)	(0,0)
s_2	(1,0)	(0,2)

Table 4.5: A game with a better pruning for agent 1.

$pr(\mathbb{G})$	t_2	t_3
$\theta([s_1])$	(1,2)	(0,0)

$pr'(\mathbb{G})$	t_2	t_3
$\theta'([s_1])$	(0, 0)	(0,2)

Table 4.6: The two prunings of Table 4.5.

This shows that the result of pruning can be made dependent not only on the agents' intentions, but also on their preferences and rationality. As we saw in Fact 4.2.8, the pruning operation is *not* in itself responsive to these characteristics[12]. To avoid empty fixed points, the agents' intentions must make up for this.

[11] The considerations about knowledge of intentions that I made in the footnote on page 69 apply with even greater force here. The epistemic analysis of non-empty pruning would have to take into account not only the agents' information about each other's intentions, but also about their picking criteria.

[12] To put it in terms I used for cleaning, I defined here an "individualistic" pruning. A more altruistic pruning would have to take not only strategies but also interdependence of picking functions into account.

The last example shows that the picking criterion is also of great importance here.

To be sure, there is a lot to be said about how various constraints on the picking functions would embody the responsiveness to others' intentions, preferences and rationality[13]. I do not, however, go in that direction here. My goal was rather to point to the fact that pruning, just like cleaning, is quite sensitive to the interactive aspects of strategic games. When transforming their strategy sets agents should take into account that they interact with other *planning* agents, that is agents who also fit their intentions sets according to what they intend. As we shall now see, this sensitivity to others has to be even more acute when pruning and cleaning can be combined.

4.3 Grouping and ruling out options

The full picture of reasoning-centered commitment of intentions certainly requires one not only to look at how pruning and cleaning transform strategic games, by also at how they *interact* with one another. To keep things relatively simple, I look at this interaction in terms of *sequential applications* of these operations. That this, I look at what happens when agents first perform one of the operations, then the other, and so on.

The first thing to notice is that cleaning and pruning do not in general commute. Table 4.7 is a counterexample, again with the elements of $\downarrow\iota_i$ indicated by the numbers in the cells. If we fix $\theta_2([t_1]) = t_1$, after one round of pruning we obtain the centre matrix $pr(\mathbb{G})$. This is in turn reduced to the rightmost matrix after one more round of cleaning. But observe that $cl(\mathbb{G}) = \mathbb{G}$ and thus that $pr(cl(\mathbb{G})) = pr(\mathbb{G}) \neq cl(pr(\mathbb{G}))$. Interestingly, in this game individualistic and altruistic admissibility give the same result: pruning does not commute with either type of cleaning.

$\mathbb{G}$	t_1	t_2
s_1		1
s_2	1, 2	1, 2

$pr(\mathbb{G})$	t_1
s_1	
s_2	1, 2

$cl(pr(\mathbb{G}))$	t_1
s_2	1, 2

Table 4.7: Counter-example to commutativity.

It should also be clear that sequences of pruning and cleaning can make strategic games empty, even when altruistic admissibility drives cleaning. The reason

[13] Observe, for instance, that in the last example both feasible outcomes after the two possible prunings are also feasible in the original game. This suggests, as I mentioned earlier (footnote on page 68), a connection between "dominated" picking functions such as θ'_1, on the one hand, and more standard game-theoretical solution concepts on the other. But observe that the connection is rather loose. Even though agent 1 is strictly worse off with the feasible outcome he gets after pruning with θ'_1, this outcome is *not* part of a dominated strategy in the original game, not even a weakly dominated one.

is that pruning can by itself make the intention sets of agents internally inconsistent. Once this stage is reached, one more round of cleaning makes the game empty, whatever notion of admissibility is running in the background. This would happen in the example of Table 4.4. One further round of cleaning eliminates all strategies for both agents.

This interaction between cleaning[14] and pruning really creates new possibilities for empty fixed points. Neither the existence of a cleaning core nor of safe outcomes is sufficient to preserve consistency of intentions. For the first case, look again at the game in Table 4.4. It has a cleaning core, namely the four combinations of s_1 and s_2 with t_1 and t_2. But, as we saw, pruning can make the intentions of all agents internally inconsistent in that game. For the second case, look at the game in Table 4.1. Here all outcomes are safe for pruning, but cleaning quickly makes this game empty.

Not even a combination of the two criteria ensures non-emptiness. Consider the game in Table 4.8, where $\theta_i([s_1]^{\mathbb{G}}_{\approx}) = s_2$. The game reached after one round of pruning is the pruning fixed point, which means that the outcomes of both (s_2, t_1) and (s_2, t_2) are safe for this operation. These two profiles are also in the cleaning core, as in fact are the two others. But by alternating the two operations, we reach the rightmost game, where agent 1 has internally inconsistent intentions.

$\mathbb{G}$	t_1	t_2
s_1	1, 2	
s_2	1	2

$pr(\mathbb{G})$	t_1	t_2
s_2	1	2

$cl(pr(\mathbb{G}))$	t_2
s_2	2

Table 4.8: A game where the combination of the two operations makes the game empty.

In all these examples, the alternation of cleaning and clustering has a *unique* fixed point[15]. But this need not be so, at least when cleaning is done with individualistic admissibility. Consider the game in Table 4.9, and assume that the picking function of agent 1 satisfies the following: $\theta_1(\{s_1, s_2\}) = s_2$, $\theta_1(\{s_1, s_2, s_3\}) = s_1$ and $\theta_1(\{s_2, s_3\}) = s_2$.

If we start by cleaning this game, only t_3 is removed. This makes all three strategies of agent 1 means-redundant in $cl(\mathbb{G})$. According to 1's picking function, only s_1 remains in $pr(cl(S_i))$, which makes t_1 inadmissible for agent 2. One more round of cleaning thus makes this game empty.

Things are different, however, if we start by pruning instead of cleaning. Only agent 1 reduces his strategy set in this case, by picking s_2 in $\{s_1, s_2\}$. This makes both t_2 and t_3 inadmissible, leaving only t_1 in $cl(pr(S_2))$. In this reduced game,

[14] For the remainder of the section I consider individualistic admissibility only. I return to altruistic admissibility in Chapter 5.

[15] Observe that this is even the case in the counterexample to commutativity (Table 4.7).

$\mathbb{G}$	t_1	t_2	t_3
s_1	1	2	
s_2	1, 2		
s_3	1		1

Table 4.9: A game with two different fixed-points.

$cl(\mathbb{G})$	t_1	t_2
s_1	1	2
s_2	1, 2	
s_3	1	

$pr(cl(\mathbb{G}))$	t_1	t_2
s_1	1	2

$cl(pr(cl(\mathbb{G})))$	t_2
s_1	2

Table 4.10: The route to the first (empty) fixed point of the game in Table 4.9.

the centre matrix in Table 4.11, s_2 and s_3 is means-redundant for 1, among which he picks s_2, leading to a non-empty fixed point.

$pr(\mathbb{G})$	t_1	t_2	t_3
s_2	1, 2		
s_3	1		1

$cl(pr(\mathbb{G}))$	t_1
s_2	1, 2
s_3	1

$pr(cl(pr(\mathbb{G})))$	t_2
s_2	1, 2

Table 4.11: The second fixed point of the game in Table 4.9.

If we ignore redundant transformations, all sequences of cleaning and pruning reach a fixed point in a finite number of steps, for every finite strategic games[16]. The example above reveals, however, that some games do not have a unique fixed point, but many different ones.

4.3.1. Definition. [Iterated transformation] Given a strategic game $\mathbb{G}$, let $t(\mathbb{G})$ be either $pr(\mathbb{G})$ or $cl(\mathbb{G})$. A *sequence of transformation of length* k is any $t^k(\mathbb{G})$ for $k \geq 0$, where $t^1(\mathbb{G}) = t(\mathbb{G})$ and $t^{k+1}(\mathbb{G}) = t(t^k(\mathbb{G}))$. A sequence of transformation $t^k(\mathbb{G})$ is a *transformation fixed point* whenever both $cl(t^k(\mathbb{G})) = t^k(\mathbb{G})$ and $pr(t^k(\mathbb{G})) = t^k(\mathbb{G})$.

As we saw in Table 4.8, ensuring a non-empty fixed point is not just a matter of looking at the intersection of the cleaning core with the set of profiles yielding outcomes that are safe for pruning. The problem is that these two notions do not take account of the possible alternation of pruning and cleaning. The following, stronger notion of safety ensures the existence of non-empty fixed points.

[16] The reference to Apt [2007], and the considerations in the footnote on page 68 are also important here.

4.3.2. Definition. [Safety for iterated transformations] The outcome x of profile $\sigma \in \Pi_{i \in I} S_i$ is:

- *Safe for iterated transformations at stage 1* whenever, for all $i \in I$:
  1. $\pi(\sigma(i)) \cap \downarrow \iota_i \neq \emptyset$.
  2. $\theta_i[\sigma(i)]^{\mathbb{G}}_{\approx} = \sigma(i)$.
- *Safe for iterated transformations at stage n + 1* whenever it is safe for iterated transformation at stage n and for all $i \in I$:
  1. $\pi^{t^n(\mathbb{G})}(\sigma(i)) \cap \downarrow \iota_i^{t^n(\mathbb{G})} \neq \emptyset$.
  2. $\theta_i[\sigma(i)]^{t^n(\mathbb{G})}_{\approx} = \sigma(i)$.
- *Safe for iterated transformations* whenever it is safe for transformation at all n.

4.3.3. Fact. [Safety for transformation and non-empty fixed points] For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking function Θ, if $\pi(\sigma)$ is safe for transformation in $\mathbb{G}$ then for all fixed points $t^{\#}(\mathbb{G})$, $\sigma \in \Pi_i t^{\#}(S_i)$.

Proof. This is shown by induction on k for an arbitrary fixed point $t^k(S_i)$. The proof is a direct application of Definition 4.3.2. ■

The presence of safe outcomes is thus sufficient to ensure that a game has no empty fixed point. In fact it ensures something stronger, namely that all fixed points have a non-empty intersection. But precisely because of that, it does not entail that any game which has no empty fixed point contains safe outcomes. If it can be shown that all games have a unique, non-empty fixed point, which is the case in all the examples considered above, then we would know that safety for transformation exactly captures non-emptiness. It remains, however, open to me whether this is the case or not.

Interestingly, the converse of Fact 4.3.3 also holds, if we impose the following constraint on picking functions.

4.3.4. Definition. [Consistent picking functions] A picking function θ_i is *consistent* if $\theta_i(X) = s_i$ whenever $\theta_i(Y) = s_i$, $X \subseteq Y$ and $s_i \in X$.

A good example of a consistent picking function is one which always picks the maximal element in some fixed ranking, for example the preference relation[17]. If s_1 is the maximal element among all strategies in X then, provided that the ranking is kept constant, s_1 stays the maximal element in all the subsets in which it appears.

[17] I draw this condition from Sen's [2002] "property α", who uses it in a decision-theoretic context as a constraint on consistency of choices.

4.3.5. Fact. [Non-empty fixed points and safe outcomes] For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking function Θ, if $\sigma \in \Pi_i t^{\#}(S_i)$ for all fixed points $t^{\#}(\mathbb{G})$, then $\pi(\sigma)$ is safe for transformation in $\mathbb{G}$.

Proof. I show by "backward" induction that $\pi(\sigma)$ is safe for transformation at any k for all sequences $t^k(\mathbb{G})$. For the basic case, take k to be the length of the longest, non-redundant fixed point of $\mathbb{G}$. I show that $\pi(\sigma)$ is safe for transformation at stage k for all sequences of that length. Observe that by the choice of k all $t^k(\mathbb{G})$ are fixed points. We thus know by assumption that $\sigma \in \Pi_{i\in I}t^k(S_i)$. But then it must be safe for transformation at stage k. If clause (1) was violated at one of these, say $t'^k(\mathbb{G})$, then we would have $cl(t'^k(\mathbb{G})) \neq t'^k(\mathbb{G})$, against the fact that $t'^k(\mathbb{G})$ is a fixed point. By the same reasoning we know that clause (2) cannot be violated either. Furthermore, by the fact that $t'^{k+1}(\mathbb{G}) = t'^k(\mathbb{G})$, we know that it is safe for transformation at all stages $l > k$.
For the induction step, take any $0 \leq n < k$ and assume that for all sequences $t^{n+1}(\mathbb{G})$ of length $n+1$, $\pi(\sigma)$ is safe for transformation at stage $n+1$. Take any $t^n(\mathbb{G})$. By our induction hypothesis, that $\pi(\sigma)$ is safe for transformation at both $cl(t^n(\mathbb{G}))$ and $pr(t^n(\mathbb{G}))$. This secures clause (2) of Definition 4.3.2, and also gives us that $\sigma \in \Pi_{i\in I}t^n(S_i)$. Now, because it is safe for transformation in $cl(t^n(\mathbb{G}))$, we know that $\pi^{cl(t^n(\mathbb{G}))}(\sigma(i)) \cap \downarrow \iota_i^{cl(t^n(\mathbb{G}))} \neq \emptyset$ for all i. But since $\pi^{cl(t^n(\mathbb{G}))}(\sigma(i)) \subseteq \pi^{t^n(\mathbb{G})}(\sigma(i))$, and the same for the intention set, we know that $\pi^{t^n(\mathbb{G})}(\sigma(i)) \cap \downarrow \iota_i^{t^n(\mathbb{G})} \neq \emptyset$ for all i. For condition (2), we also know that $\theta_i[\sigma(i)]_{\approx}^{cl(t^n(\mathbb{G}))} = \sigma(i)$ for all i from the fact that $\pi(\sigma)$ is safe for transformation at stage $n+1$. By Lemma 4.3.6 (below) and the assumption that θ_i is consistent for all i, we can conclude that $\theta_i[\sigma(i)]_{\approx}^{t^n(\mathbb{G})} = \sigma(i)$, which completes the proof because we took an arbitrary $t^n(\mathbb{G})$. ■

4.3.6. Lemma. *For any game strategic game $\mathbb{G}$ and intention set ι_i and strategy $s_i \in cl(S_i)$, $[s_i]_{\approx}^{\mathbb{G}} \subseteq [s_i]_{\approx}^{cl(\mathbb{G})}$.*

Proof. Take any $s'_i \in [s_i]_{\approx}^{\mathbb{G}}$. Since $s_i \in cl(S_i)$, we know that there is a $\sigma_{j\neq i}$ such that $\pi(s_i, \sigma_{j\neq i}) \in \downarrow\iota_i$. But because $s'_i \approx s_i$, it must also be that $\pi(s'_i, \sigma_{j\neq i}) \in \downarrow\iota_i$, and so that $s'_i \in cl(S_i)$. Now, observe that $\{\sigma \in \Pi_{i\in I}cl(S_i) : \sigma(i) = s_i\} \subseteq \{\sigma \in S_i : \sigma(i) = s_i\}$, and the same for s'_i. But then, because $s'_i \approx s_i$, it must also be that $s'_i \in [s_i]_{\approx}^{cl(\mathbb{G})}$. ■

From the last two Facts we obtain, as a direct corollary, that if all players intend safe outcomes then no fixed-point is empty, and we can "track" safe outcomes in the agents' original intentions by looking at those they keep intending in all fixed-points.

4.3.7. Corollary. *For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking function Θ, the following are equivalent.*

1. *For all i there is a $\pi(\sigma)$ in $\downarrow\iota_i$ that is safe for transformation in $\mathbb{G}$.*

2. *$\pi(\sigma) \in\downarrow\iota_i^{t^{\#}(\mathbb{G})}$ for all fixed-points $t^{\#}(\mathbb{G})$.*

The existence of empty transformation fixed points once again shows the importance of taking each others' intention into account while simplifying decision problems. To be sure, the pruning and cleaning *do* commute when there is only one agent.

4.3.8. Fact. $pr(cl(\mathbb{G})) = cl(pr(\mathbb{G}))$ for any strategic game $\mathbb{G}$ with only one agent, intention set ι_i and picking function θ_i.

Proof. This is a direct consequence of the following lemma[18].

4.3.9. Lemma. *For any strategic game $\mathbb{G}$ and intention set ι_i, if $s_i \not\in cl(S_i)$ with individualistic admissibility then $\theta_i([s_i]^{\mathbb{G}}_{\approx})^{pr(\mathbb{G})} \not\in cl(pr(\mathbb{G}))$. The converse also holds for all strategic games $\mathbb{G}$ with one agent.*

For the first part, assume that $s_i \not\in cl(S_i)$. This means that for all profiles, $\pi(s_i) \cap \downarrow\iota_i = \emptyset$. This means, in turn, that $\pi(s'_i) \cap \downarrow\iota_i = \emptyset$ for all s'_i such that $s'_i \in [s_i]^{\mathbb{G}}_{\approx}$. So, whatever is the value of $\theta_i([s_i]^{\mathbb{G}}_{\approx})$, we know that $\pi^{pr(\mathbb{G})}(\theta_i([s_i])) \cap \iota_i^{pr(\mathbb{G})} = \emptyset$, and so that $\theta_i([s_i]^{\mathbb{G}}_{\approx})^{pr(\mathbb{G})} \not\in cl(pr(\mathbb{G})$. The second part follows the same line, this time using the fact that in single agent cases, $\pi(s_i)$ is a singleton. ■

4.4 Conclusion

The considerations at the end the last section reflect the overall concern of this chapter. In genuine strategic situations the reasoning-centered commitment of intention takes on a crucial interactive dimension. We saw in particular how important it is for agents who rule out inadmissible options and group redundant alternatives to take into account the fact that they interact with other agents who similarly transform their decision problems. If they fail to take the others into account, there might simply be no more options for them to decide.

To stress the matter a little more, it is worth recalling that when there is only one decision maker, neither cleaning nor pruning and not even a combination of the two can eliminate all options in a decision problem, if the agent has realizable intentions to start with. What is more, in all these cases a single application of these transformations is enough to reach a fixed point, and as we have just seen, the order of these two transformations is not important. In other words, the interesting complications that we faced in this chapter genuinely stemmed from the interaction of planning agents.

[18] The matrix in Table 4.7 shows that the "furthermore" really only holds for single-agent strategic games.

This would become even more clear by supplementing the analysis of this chapter with an epistemic dimension. Intuitively, agents can only take care of the intentions and picking criteria of others if they *know* what these are. As we saw in the previous chapter, this is precisely the kind of claim that can be given a formal treatment using epistemic models. In the next chapter I take a step in that direction, by looking at how epistemic models of games with intentions evolve in the process of cleaning. This will, however, only be a first step since, for instance, I do not look at pruning. But putting together interactive information and cleaning of decision problems will unveil yet another dimension of intention-based practical reasoning, in particular new "epistemically laden" admissibility criteria. What is more, the "logical" methods that I use allow for an explicit representation of practical *reasoning* in games with intentions, something which has so far been much discussed but still left implicit.

I should also stress once again that looking at intention revision is another way to cope with the fact that transformation of decision problems can lead to empty games. I did not look at intention revision here, because it would have lead us too far along a tangent to reasoning-centered commitment. Its full-fledged analysis requires one to draw from the theory of *belief* revision—as e.g. in Rott [2001]—as well as from the in depth analysis of Bratman [1987, chap.5] and van der Hoek et al. [2007]. But the considerations in this chapter could also be used as inputs to such a theory of intention revision. Intentions which cannot become internally inconsistent after some game transformations are obviously good things to adopt in the need of revision. In other words, the conditions that I isolated in this chapter could arguably be turned into "fall back" ones, thus contributing to the understanding of intention revision.

Chapter 5

Logics for practical reasoning with intentions

In the previous chapters I used formal models of rational interaction to deepen our understanding of three important facets of practical reasoning with intentions: the reasoning-centered and volitive commitments of intentions, the information in games and the rationality of the agents. In this chapter I use *dynamic epistemic logic* to combine these aspects into a unified theory of practical reasoning of planning agents in interactive situations. This will give us a better understanding of how the information about intentions, preferences and mutual knowledge becomes involved in practical reasoning, and also how it changes in the course of this process. Looking at these games with intentions through the lenses of logic also provides a concrete representation of practical *reasoning.* Such formal languages come with well-known proof systems, in which inferences involving intentions, knowledge, preferences and actions are actually worked out.

To get us acquainted with the logical "toolbox" that I use throughout the chapter, in Section 5.1 I look at simple preference structures. I then turn to epistemic models for games with intentions (Section 5.2). As we shall see, these models express themselves naturally through logical languages, and they have a lot to say about the relation between intention, information, and rational agency. In Section 5.3 dynamic epistemic logic comes into play in order to capture transformations of game models. I show that it unveils natural epistemic variants of the cleaning operation, that it allows for a more systematic study of intention overlap and of conditions under which cleaning is "enabled".

5.1 Preliminaries: modal logic for preferences

All the decision problems I have studied so far included a representation of the agents' preferences. They were described in very similar terms and they shared some common properties. In this section I highlight these properties and, at the same time, deploy most of the logical machinery that I use in this chapter.

The study of preferences from a logical point of view, the so-called "preference logic", has a long history in philosophy (see e.g. von Wright [1963] and Hansson [2001]) and computer science (see e.g. Boutilier [1994] and Halpern [1997]). The logic I present here has been developed by Johan van Benthem, Patrick Girard, Sieuwert van Otterloo and myself during recent years. The reader can consult [van Benthem et al., 2005; van Benthem et al., Forthcoming] for more details.

5.1.1 Preference models and language

All the logical languages I use in this chapter have been devised to talk about classes of structures, generally classes of *relational frames*[1]. These are simply sets of states interconnected by a number of relations. The extensive and strategic decision problems of Chapter 2, the strategic games and the epistemic models of Chapter 3 can all be seen as relational frames. To study the features of preferences in abstraction from their representations into some particular games or decision problems is just to look at the preference component of these frames.

5.1.1. Definition. [Preference frames] A *preference frame* $\mathbb{F}$ is a pair $\langle W, \succeq \rangle$ where:

- W is a non-empty set of states,
- $\succeq$ is a reflexive and transitive relation, i.e. a "preorder", over W. Its strict subrelation, noted $\succ$, is defined as $w \succ w'$ iff $w \succeq w'$ but $w' \not\succeq w$.

The relation $w \succeq w'$ should be read "w is *at least as good as* w'." In all the models of the previous chapters this relation was assumed to be reflexive, transitive and, most of the time, total. The relation of strict preference $\succ$ is the irreflexive and transitive sub-relation of $\succeq$. The reader can check that these properties indeed follow from the definition of $\succ$. If $w \succ w'$, we say that w is strictly preferred to w'.

Adopting a logical point of view on preference frames—in fact, on any class of relational frames—means talking about them in terms of some formal language. In this chapter I use *propositional modal languages*, which are essentially propositional Boolean languages supplemented with modal operators, in order to talk about the properties of the relation. The language for preference frames is defined as follows.

[1] I borrow this terminology, like almost all the definitions and techniques used in this section, from Blackburn et al. [2001].

5.1.2. Definition. [Preference language] Given a set of atomic proposition prop, the language $\mathcal{L}_{\mathcal{P}}$ is inductively defined as follows[2].

$$\phi ::= p \mid \phi \wedge \phi \mid \neg \phi \mid \Diamond^{\leq}\phi \mid \Diamond^{<}\phi \mid E\phi$$

The "boolean" fragment of this language thus contains the propositions together with the conjunction and negation operators. I use $\top$ to abbreviate the tautology $p \rightarrow p$ and $\bot$ to abbreviate $\neg\top$. The modal operators are $\Diamond^{\leq}$, $\Diamond^{<}$ and E. Formulas of the form $\Diamond^{\leq}\phi$ and $\Diamond^{<}\phi$ should be read, respectively, as "ϕ is true in a state that is considered at least as good as the current state" and "ϕ is true in a state that is considered strictly better than the current state". $E\phi$ is a "global" modality. It says that "there is a state where ϕ is true". As usual in modal logic, I take $\Box^{\leq}\phi$ to abbreviate $\neg\Diamond^{\leq}\neg\phi$. This formula can be read as "ϕ holds in all states that are at least as good as the current one". $\Box^{<}\phi$ is defined similarly. $A\phi$, which abbreviates $\neg E \neg \phi$, is taken to mean "ϕ holds in all states".

The key step in any logical investigation of a certain class of frames is to connect the formulas of the language with elements of the frames. This is done by defining a *model*, which is essentially an assignment of truth values to the propositions in prop, and the *truth conditions* for the other formulas of the language.

5.1.3. Definition. [Preference models] A *preference model* $\mathbb{M}$ is a preference frame $\mathbb{F}$ together with a *valuation function* V : prop $\rightarrow \mathcal{P}(W)$ that assigns to each propositional atom the set of states where it is true. A *pointed preference model* is a pair $\mathbb{M}, w$.

5.1.4. Definition. [Truth and validity in $\mathcal{L}_{\mathcal{P}}$]

$$\begin{array}{lll}
\mathbb{M}, w \models p & \text{iff} & w \in V(p) \\
\mathbb{M}, w \models \phi \wedge \psi & \text{iff} & \mathbb{M}, w \models \phi \text{ and } \mathbb{M}, w \models \psi \\
\mathbb{M}, w \models \neg\phi & \text{iff} & \mathbb{M}, w \not\models \phi \\
\mathbb{M}, w \models \Diamond^{\leq}\phi & \text{iff} & \text{there is a } v \text{ such that } v \succeq w \text{ and } \mathbb{M}, v \models \phi \\
\mathbb{M}, w \models \Diamond^{<}\phi & \text{iff} & \text{there is a } v \text{ such that } v \succ w \text{ and } \mathbb{M}, v \models \phi \\
\mathbb{M}, w \models E\phi & \text{iff} & \text{there is a } v \text{ such that } \mathbb{M}, v \models \phi
\end{array}$$

A formula ϕ is *valid in a preference model* $\mathbb{M}$, denoted $\mathbb{M} \models \phi$, whenever $\mathbb{M}, w \models \phi$ for all $w \in W$. A formula is *valid in a preference frame* $\mathbb{F}$ whenever it is valid in

[2] A few guidelines for the reader unaccustomed to this way of defining logical languages. $\phi ::= p \mid \phi \wedge \phi \mid ...$ means that a formula ϕ of that language is either a proposition letter p from prop, a conjunction of formulas of the language, the negation of a formula of the language, and so on. I do not make any assumption regarding the finiteness of prop. In most of what follows I use a multi-agents version of this language, in which I index the modalities with members of a set I of agents. I omit this here, but the results about the preference language generalize naturally to the multi-agent case.

all preference models $\mathbb{M} = \langle \mathbb{F}, V \rangle$. Finally, a formula is valid in a class of models M whenever it is valid in all models $\mathbb{M} \in \mathsf{M}$. Validity with respect to classes of frames is defined in the same way.

These truth conditions are intended to capture the intuitive meaning of the various connectives just described. For example, the truth condition for $\Diamond^{\leq}\phi$ literally states "ϕ is true in a state that is considered at least as good as the current state."

Equipped with a language and an interpretation for it, we can start the logical investigation. In this section and the subsequent ones, it divides into two main inquiries.

First I look at what can and what cannot be *said* about a given class of frames with the language at hand. This is called looking at the *expressive power*. In the case of preference frames, we shall see that some properties of the relations $\succeq$ and $\succ$ find clear expression in $\mathcal{L}_{\mathcal{P}}$, while others are beyond its reach. Furthermore, in this language one can unveil features of the intended class of frames that are not obvious at first sight. In the case of preference frames, we shall see that we can study in $\mathcal{L}_{\mathcal{P}}$ properties of "lifted" preference relations, from preference between states to preference between sets of states.

Beside questions related to expressive power, most logical investigations look at what kind of *inferences* can be made about some class of frames. This is done by providing a proof system, i.e. a *logic*, in which one can derive formulas that are true or valid with respect to some (classes of) frames. Here I use *axiomatic proof systems*, which neatly encapsulate key properties of the class of frames we want to talk about.

5.1.2 Expressive power

To show that a property of a certain class of frames is expressible in given language, one has to provide a formula that is valid in a class of frames exactly when all the frames in that class have that property. More precisely, one has to find a formula ϕ such that ϕ is valid in a class of frame F iff all the frames in F have this property. If we can find such a formula, we say that we have a *correspondent* for that property in the language.

Transitivity and reflexivity of $\succeq$ have well-known correspondent in $\mathcal{L}_{\mathcal{P}}$[3] :

$$\Diamond^{\leq}\Diamond^{\leq}\phi \rightarrow \Diamond^{\leq}\phi \qquad \text{(Transitivity)}$$

$$\phi \rightarrow \Diamond^{\leq}\phi \qquad \text{(Reflexivity)}$$

Totality is also expressible, but its corresponding formula crucially uses the global modality E.

5.1.5. FACT. The following corresponds to $\succeq$ being a total relation.

$$\phi \wedge E\psi \rightarrow (\Diamond^{\leq}\psi \vee E(\psi \wedge \Diamond^{\leq}\phi)) \qquad \text{(Totality)}$$

[3] The correspondence arguments are well know. See again Blackburn et al. [2001, chap.3].

Proof. It is easy to see that this formula is valid on a preference frame, provided its relation $\succeq$ is total. For the other direction, take a preference model $\mathbb{M}$ where this formula is valid, which contains only two states w and w' and where ϕ and ψ are true only at w and w', respectively. The truth condition for E gives $\mathbb{M}, w \models \phi \wedge E\psi$, and so the consequent of (Totality) must also be true there. But then either $\mathbb{M}, w \models \Diamond^{\leq}\psi$, which means that $w' \succeq w$ or $\mathbb{M}, w \models E(\psi \wedge \Diamond^{\leq}\phi)$, which means, by the way we devised our model, that $w \succeq w'$. ■

The properties of $\succ$ are harder to express in $\mathcal{L}_{\mathcal{P}}$. One can easily say that it is a sub-relation of $\succeq$ with the following:

$$\Diamond^{<}\phi \rightarrow \Diamond^{\leq}\phi \qquad \text{(Inclusion)}$$

It is, however, more intricate to ensure that it is precisely the sub-relation that I defined in 5.1.1. In particular, irreflexivity of $\succ$ is *not* expressible in $\mathcal{L}_{\mathcal{P}}$. That is, there is no formula of $\mathcal{L}_{\mathcal{P}}$ that is valid on a class of frames if and only if $\succ$ is irreflexive in all frames of this class. To show this requires a notion of *invariance* between preference frames or models.

The best known is probably that of *bisimulation*[4]. Two pointed models $\mathbb{M}, w$ and $\mathbb{M}', v$ are bisimilar when, first, they make the same propositions true and, second, if there is a w' such that $w' \succeq w$, one can find a v' bisimilar to w' such that $v' \succeq' v$, and vice-versa from $\mathbb{M}'$ to $\mathbb{M}$. A standard modal logic argument shows that if two pointed preference models are bisimilar, then they make exactly the same formulas of $\mathcal{L}_{\mathcal{P}}$ true.

With this in hand, to show that a given property is not definable in $\mathcal{L}_{\mathcal{P}}$ boils down to finding two bisimilar pointed models, one that does and the other that does not have the property. With such an argument one can show that irreflexivity is not definable[5].

The various properties of $\succeq$ and $\succ$ thus provide benchmarks to assess the expressive capacities of $\mathcal{L}_{\mathcal{P}}$. But, as I mentioned, the real interest of such a language is that it can capture in a perspicuous manner features of preference frames that would otherwise be quite intricate to grasp.

A good example is the "lifting" of $\succeq$ and $\succ$, which are relations between states, to relations between sets of states. One might consider that, for example, a set of states Y is "at least as good" as a set of state X whenever for all states in X one can find a state that is at least as good in Y. One can easily capture this "lifted" preference relation with binary preference *statements* between formulas of $\mathcal{L}_{\mathcal{P}}$. After all, formulas neatly correspond to sets of states in a preference model, namely the sets of states where they are true.

[4] Here I simply sketch the definition of this notion. The precise definition can be found in Appendix 5.5.

[5] By an argument similar (and in fact related) to the one for inexpressibility of irreflexivity one can also show that the modality $\Diamond^{<}$ is not definable in terms of $\Diamond^{\leq}$. In other words, to talk directly about the $\succ$ relation one has to introduce a separate modality.

$$\phi \leq_{\forall\exists} \psi \quad \Leftrightarrow \quad A(\phi \rightarrow \Diamond^{\leq}\psi) \qquad \text{(Lifted relation)}$$

The reader can check that the formula $\phi \leq_{\forall\exists} \psi$ does indeed correspond to the fact that for all the states where ϕ is true one can find a state that is at least as good where ψ is true. In other words, this formula expresses that ψ is at least as good as ϕ.

Are properties of $\succeq$ also lifted to $\leq_{\forall\exists}$? For example, does it follow from the fact that $\succeq$ is total that $\leq_{\forall\exists}$ is also a total relation between sets of states? This is not so obvious merely from an examination of preference models, but it becomes quite transparent if one goes through the truth conditions of $\phi \leq_{\forall\exists} \psi$.

5.1.6. Fact. With respect to the class of preference models, if $\succeq$ is total then for all formulas ϕ and ψ of $\mathcal{L}_{\mathcal{P}}$, either $\phi \leq_{\forall\exists} \psi$ or $\psi \leq_{\forall\exists} \phi$.

Proof. Take a preference model $\mathbb{M}$ where $\succeq$ is total, and two formula ϕ and ψ. If either ϕ or ψ is not satisfied in $\mathbb{M}$, then we are done. Assume then that both are satisfied, and take any state w such that $\mathbb{M}, w \models \phi$. We have to show that there is a w' such that $w' \succeq w$ and $\mathbb{M}, w' \models \psi$. Assume this is not the case, i.e. that for all w' such that $w' \succeq w$, $w' \not\models \psi$. Given totality of $\succeq$, and the fact that ψ is satisfied, this means that for all w'' such that $\mathbb{M}, w'' \models \psi$, $w \succ w''$. But this is enough to show that $\mathbb{M} \models \psi \leq_{\forall\exists} \phi$. ■

This kind of analysis can be carried further to other properties of $\leq_{\forall\exists}$, such as reflexivity and transitivity, and even to alternative lifted relations. I shall not pursue this further here[6]. For now it is enough to know that by devising a modal language to talk about a given class of frames one can express in a very clear way notions that would otherwise have been rather opaque. For now, I want to turn briefly to the second side of logical inquiry, namely inferences and proof systems.

5.1.3 Axiomatization

One can precisely capture reasoning about a given class of frames by providing a system of *axioms* and *inference rules* such that all formulas valid on that class of frame are derivable in that system. This is called showing *completeness* of an axiom system. This is usually more difficult than showing that the system is *sound*, i.e. that everything that can be derived in it is a valid formula. If we can show both, then we know that the set of valid formulas is exactly the set of formulas that are derivable in that system.

There is a sound and complete axiom system for the class of preference frames. It can be found on page 90. The reader will recognize in this table many formulas

[6] The reader can find various other definability and lifting results in the Appendix 5.5.2 and in Liu [2008].

that we have already encountered in this section. This is, of course, no coincidence[7]. What can be deduced using a given language about a given class of frames depends on its expressive power. This is, in a way, what the following theorem says.

5.1.7. THEOREM. *The logic $\Lambda^{\mathcal{L}_\mathcal{P}}$ is sound and complete with respect to the class of preference models. With (Tot) it is sound and complete with respect to the class of total preference models.*

Proof. See van Benthem et al. [Forthcoming]. ■

This is where I stop this brief investigation into the logic of abstract preference frames. In the next sections I follow essentially the same methodology: I define logical languages for the intended classes of frames and examine what can be said and which sorts of reasoning can be conducted with them. For the class of preference frames, this methodology has already paid off: it has shed light on properties of preference relation between sets of states. But the usefulness of a logical point of view for intention-based practical reasoning really reveals itself on more complex models, to which I now turn.

5.2 Logic for games with intentions

In this section I take a closer (logical) look at the epistemic models of games with intentions that I used in Chapter 3. We shall see that logical methods shed new light on how planning agents use the information they have about each others' information and intentions to reason in strategic interaction.

5.2.1 Language for epistemic game models with intentions

In Chapter 2 and 3 I took care to distinguish between strategy profiles and outcomes in the representation of decision problems. This provided an encompassing point of view, bringing under the same umbrella strategic games and models of decision making under uncertainty. In the present chapter, however, I ignore this distinction between outcomes and profiles in order to simplify the analysis.

Recall that in Chapter 3 *epistemic models* were always constructed on the basis of a strategic game $\mathbb{G}$. Here I directly define *epistemic game frame*, packing both the game and the epistemic information into a single structure. I nevertheless always assume that some strategic game $\mathbb{G}$ can be read off from any epistemic game frame. This will simplify the analysis, without any great loss of generality.

[7] Modulo certain restrictions on the shape of the formulas, there is a tight connection between them corresponding to a given property and, so to speak, "axiomatizing" it. For more details about this phenomenon, which is called Sahlqvist correspondence, see [Blackburn et al., 2001, p.157-178].

- All propositional tautologies.

- $S4$ for $\Diamond^{\leq}$:
  (K) $\Box^{\leq}(\phi \wedge \psi) \leftrightarrow \Box^{\leq}\phi \wedge \Box^{\leq}\psi$
  (Trans) $\Diamond^{\leq}\Diamond^{\leq}\phi \rightarrow \Diamond^{\leq}\phi$
  (Ref) $\phi \rightarrow \Diamond^{\leq}\phi$
  (Tot) $\phi \wedge E\psi \rightarrow (\Diamond^{\leq}\psi \vee E(\psi \wedge \Diamond^{\leq}\phi))$

- For $\Diamond^{<}$:
  (K) $\Box^{<}(\phi \wedge \psi) \leftrightarrow \Box^{<}\phi \wedge \Box^{\leq}\psi$

- $S5$ for E:
  (K) $A(\phi \wedge \psi) \leftrightarrow A\phi \wedge A\psi$
  (Trans) $EE\phi \rightarrow E\phi$
  (Ref) $\phi \rightarrow E\phi$
  (Sym) $E\phi \rightarrow AE\psi$

- Interaction axioms.
  Inc$_1$ $\Diamond^{<}\phi \rightarrow \Diamond^{\leq}\phi$
  Inc$_2$ $\Diamond^{\leq}\phi \rightarrow E\phi$
  Int$_1$ $\Diamond^{\leq}\Diamond^{<}\phi \rightarrow \Diamond^{<}\phi$
  Int$_2$ $\phi \wedge \Diamond^{\leq}\psi \rightarrow (\Diamond^{<}\psi \vee \Diamond^{\leq}(\psi \wedge \Diamond^{\leq}\phi))$
  Int$_3$ $\Diamond^{<}\Diamond^{\leq}\phi \rightarrow \Diamond^{<}\phi$

- The following inference rules:

  Nec If ϕ is derived then infer $\Box^{\leq}\phi$. Similarly for $\Box^{<}$ and A.

Table 5.1: The axiom system for $\Lambda_{\mathcal{L}_{\mathcal{P}}}$.

5.2.1. Definition. [Epistemic game frames with intentions] An *epistemic game frame with intentions* $\mathbb{G}$ is a tuple $\langle I, W, \{\iota_i, \sim_i, \succeq_i\}_{\in I}\rangle$ such that

- I is a finite set of agents.

- W is a finite set of states, viewed as strategy profiles. For convenience I keep the notation $w(i)$ for the i^{th} component of w.

- $\iota_i : W \rightarrow \mathcal{P}(\mathcal{P}(W))$ is a function that assigns to each state the intention set of i at that state. For each w and i, $\iota_i(w)$ is a filter and does not contain the empty set.

- $\sim_i$ is an equivalence relation on W such that if $w \sim_i w'$ then $w_i = w'_i$ and $\iota_i(w) = \iota_i(w')$. As in Chapter 3, $[w]_i$ denotes $\{w' : w \sim_i w'\}$.

- $\succeq_i$ is a total, reflexive and transitive preference relation on W.

There is indeed much similarity between these frames and the game models I used in Chapter 3. Instead of a general assignment of strategy and intention profiles to abstract states, I use a set of profiles W to which are assigned intention sets. Just as in the work of van Benthem [2003], strategy profiles act here directly as states.

This modelling decision is of course "logically" driven. I want to build a relational frame within which I will interpret a modal language. The reader will have recognized the preference relation $\succeq_i$ from the previous section, and the epistemic accessibility relation $\sim_i$ from Chapter 3. The latter is constrained exactly as before. Agents are assumed to know, at each state, their strategy choices and their intentions. Thus the condition that if $w \sim_i w'$ then $w(i) = w'(i)$ and $\iota_i(w) = \iota_i(w')$.

The intention function ι_i is specified as in Chapter 3. In logical vocabulary, it is a *neighbourhood function* which returns, for each state, the intention set of each agent at that state. I assume directly that these intentions are internally consistent, agglomerative and closed under superset. Recall that this means that the intention sets are *consistent filters* . As I mentioned in Chapter 2, this greatly simplifies the analysis. I shall show how in a moment, after the introduction of the language and its semantics.

As in the previous section, this language is a modal one, with the exception that it includes "constants"—σ, σ' and so on—which directly refer to strategy profiles in epistemic game frames. These constants are known as *nominals* in the modal logic literature, and languages that contain them are called *hybrid*[8].

5.2.2. Definition. [Language for epistemic game frames] Given a set of atomic propositions PROP and a set of nominals S, let $\mathcal{L}_{\mathcal{GF}}$ be the language defined as:

$$\phi ::= p \mid \sigma \mid \phi \wedge \phi \mid \neg\phi \mid \Diamond^{\leq}\phi \mid \Diamond^{<}\phi \mid K_i\phi \mid I_i\phi \mid E\phi$$

We are now familiar with the preference fragment of that language. Formulas of the form $K_i\phi$ should be read "i knows that ϕ" and those of the form $I_i\phi$ as "i intends that ϕ." As in the previous section, these connectives have duals. For $\neg K_i \neg\phi$ I use $\Diamond_i\phi$, which means "i considers ϕ possible", and for $\neg I_i \neg \phi$ I use $\mathrm{i}_i\phi$, meaning "ϕ is compatible i's intentions."

Models for epistemic game frames are essentially devised as in the previous section, with especial care for the valuation of nominals.

5.2.3. Definition. [Models for epistemic game frames] A *model* $\mathbb{M}$ is an epistemic game frame $\mathbb{G}$ together with a *valuation function* $V : (\text{PROP} \cup S) \to \mathcal{P}(W)$ that assigns to each propositional atom and nominal the set of states where it is true, with the condition that for all $\sigma \in S$, $V(\sigma)$ is a singleton. A *pointed game model* is a pair $\mathbb{M}, w$.

[8] Key references on hybrid logic are Blackburn et al. [2001, chap.7] and ten Cate [2005].

5.2.4. Definition. [Truth in $\mathcal{L}_{\mathcal{GF}}$] Formulas of the form $\Diamond_i^{\leq}\phi$, $\Diamond_i^{<}\phi$ and $E\phi$ are interpreted as in 5.2.4.

$$\begin{array}{lll}
\mathbb{M}, w \models \sigma & \text{iff} & w \in V(\sigma). \\
\mathbb{M}, w \models K_i\phi & \text{iff} & \text{for all } w' \text{ such that } w \sim_i w',\ \mathbb{M}, w' \models \phi. \\
\mathbb{M}, w \models I_i\phi & \text{iff} & ||\phi|| \in \iota_i(w), \text{ where } ||\phi|| = \{w' : \mathbb{M}, w' \models \phi\}.
\end{array}$$

The nominals are essentially interpreted in the same fashion as atomic propositions. It is the special clause on the valuation function V that turns them into real "names" for strategy profiles. The knowledge operator K_i is interpreted as in standard epistemic logic[9].

The interpretation of $I_i\phi$ is adapted to the fact that ι_i is a neighbourhood function. $I_i\phi$ is true at a state w if and only if i intends that ϕ at that state, i.e. if and only if the interpretation of ϕ is in the intention set of i at w.

Here the assumption that ι_i is a filter becomes very useful. It allows one almost to forget about the structure of a given neighbourhood $\iota_i(w)$ and look only at its "core", $\bigcap_{X \in \iota_i(w)} X$, which I once again denote $\downarrow\iota_i(w)$. By this I mean that, instead of saying "agent i intends to achieve ϕ at w" if and only if $||\phi|| \in \iota_i(w)$, one can just say that "agent i intends to achieve ϕ at w" if and only if ϕ holds at all $w' \in \downarrow\iota_i(w)$. Indeed, if the later is the case then we know that $\downarrow\iota_i(w) \subseteq ||\phi||$, and since $\iota_i(w)$ is closed under supersets, we also know that $||\phi|| \in \iota_i(w)$. The other direction is a direct consequence of closure under intersection of $\iota_i(w)$ and the finiteness assumption on W.

To say that agent i intends to achieve ϕ at w if and only if ϕ holds at all $w' \in \downarrow\iota_i(w)$ indeed reminds of the "relational" definition for the knowledge operator K_i. This is no coincidence. When neighbourhoods are filters, there is a straightforward and general back-and-forth correspondence between them and a more classical relational semantic[10]. Here I stick to the neighbourhood approach for two reasons. First, it permits one easily to drop assumptions on ι_i, for example the closure under supersets, if one finds them counterintuitive. The axiomatization result of Section 5.2.3, for example, can be easily adapted to classes of frames where ι_i is less constrained. In other words, the neighbourhood approach allows for a greater generality. The approach also allows a more intuitive presentation of the simple intention revision policy that I use in Section 5.3.1. Throughout the chapter, however, I often use this correspondence either in the formal results or to fix intuitions.

We can in fact already profit from it to understand the truth conditions of the dual $\mathsf{i}_i\phi$. In the general case, i.e. when the neighbourhood are not necessarily filters, one has to include a separate clause to ensure that $\mathsf{i}_i\phi$ really is interpreted

[9] Recall the references on the topic in the Introduction, Section 1.3.

[10] The correspondence is obvious for finite frames, as the argument in the previous paragraph shows. In the general case, neighbourhood functions such as ι_i define Alexandroff topologies, from which there is a direct translation into relational frames. See Chellas [1980], Aiello et al. [2003] and Pacuit [2007] for the details of this correspondence.

as "ϕ is compatible with i's intentions". One then requires that $\mathbb{M}, w \models \mathsf{i}_i\phi$ iff $W - ||\phi|| \notin \iota_i(w)$. But given that $\iota_i(w)$ is a filter, we automatically get that $\mathbb{M}, w \models \mathsf{i}_i\phi$ iff there is a $w' \in \downarrow\iota_i(w)$ such that $\mathbb{M}, w' \models \phi$. Hence, by closure under supersets, we know that w' is in all $X \in \iota_i(w)$, that is in all of i's intentions at w. This clearly, and in a much more intuitive manner than with the neighbourhood definition, boils down to saying that $\mathsf{i}_i\phi$ is true whenever ϕ is compatible with i's intentions.

I am now ready to put $\mathcal{L}_{\mathcal{GF}}$ to use on epistemic game frames. As in the previous section, I look first at what it can say about these frames, and then look at what kind of reasoning can be done with it.

5.2.2 Expressive power

I now use the expressive power of $\mathcal{L}_{\mathcal{GF}}$ to investigate more systematically conditions on the information, intentions, strategy choices and preferences of agents in epistemic game frames. In particular, I show that knowledge of one's own strategy choice and intentions bears unexpected consequences for the intention-related rationality constraints that I used in the previous chapters. Furthermore, we shall see that one can read off various knowledge-based conditions for Nash equilibria from epistemic game frames, using $\mathcal{L}_{\mathcal{GF}}$. This provides a connection between well-known results in epistemic foundations of game theory and shows that the current framework is quite broadly encompassing.

Definition 5.2.1 imposes three conditions on what the agents intend, know and choose. I present them briefly before discussing stronger conditions. First, agents are assumed to know their own strategy choices. To spell out the correspondent of this notion, I need to express the notion of strategy choice itself. This crucially uses the expressive power provided by nominals. Once we have it, the correspondent of knowledge of strategy choice is quite obvious[11].

$$s_i \Leftrightarrow \bigvee_{\sigma(i)=s_i} \sigma \qquad \text{(i plays strategy } s_i\text{)}$$

$$s_i \rightarrow K_i s_i \qquad \text{(Knowledge of strategy choice)}$$

The argument for the correspondence for the second formula, the knowledge of strategy choices, is relatively straightforward. More interestingly, one can see in Figure 5.1 what this condition boils down to. For all states w it makes the set $[w]_i$ completely included in the set of states where the agent plays the same strategy, i.e. the set of w' such that $w'(i) = w(i)$.

The condition that the intention sets $\iota_i(w)$ are consistent filters has two well-known correspondents in $\mathcal{L}_{\mathcal{GF}}$. On the one hand, a standard argument in neighbourhood semantic shows that ι_i is closed under conjunction and disjunction, so

[11] In the following I slightly abuse notation and write $\sigma(i)$ instead of $V(\sigma)(i)$.

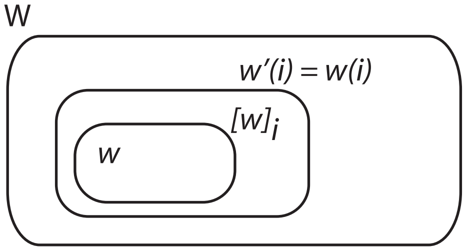

Figure 5.1: The graphical representation of knowledge of strategy choice.

that it is a filter, whenever the following hold[12].

$$I_i(\phi \wedge \psi) \leftrightarrow I_i\phi \wedge I_i\psi \qquad \text{(Intention closure)}$$

For consistency, we have the following.

$$\mathsf{i}_i\top \qquad \text{(Intention consistency)}$$

Here, once again, the fact that we can see $\mathsf{i}_i\phi$ as true at w whenever there is a $w' \in \downarrow\iota_i(w)$ such that ϕ holds at w' makes the correspondence argument completely straightforward. Indeed, since $\top$ is true at all states, to ask for the validity of $\mathsf{i}_i\top$ boils down to requiring that there is at least one state w' in $\iota_i(w)$, for all w.

Definition 5.2.1 also imposes that the agents know what they intend, i.e. that in all states that they consider possible, they have the same intentions. This can be illustrated as in Figure 5.2, and translates in $\mathcal{L}_{\mathcal{GF}}$ as follows.

5.2.5. Fact. The following corresponds to the fact that for all w, w', if $w' \sim_i w$ then $\iota_i(w') = \iota_i(w)$.

$$I_i\phi \rightarrow K_iI_i\phi \qquad \text{(Knowledge of Intention)}$$

Proof. The right-to-left direction is obvious. For left to right, take a model $\mathbb{M}$ with two states, w and w' such that $w \sim_i w'$, where the formula is valid. Fix $\downarrow\iota_i(w) = \{w'\}$ and make ϕ true only at w'. By definition, we get that $\mathbb{M}, w \models I_i\phi$, and thus that $\mathbb{M}, w \models K_iI_i\phi$. This means that $\mathbb{M}, w' \models I_i\phi$. But by the way we fixed the valuation of ϕ, it has to be that $\downarrow\iota_i(w') = \{w'\}$, which means that $\iota_i(w) = \iota_i(w')$. ■

[12] The right-to-left direction of the biconditional ensures agglomerativity, while left-to-right ensures closure under supersets. Thus, if one wants to drop one of these two assumption it is enough to look at models where only one direction holds. See Pacuit [2007] for the detailed argument.

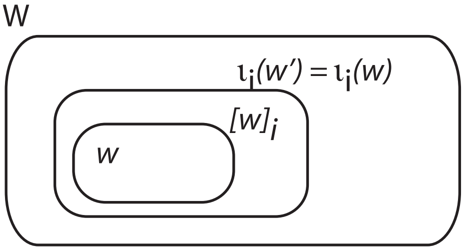

Figure 5.2: Knowledge of intentions ensures that $[w]_i \subseteq \{w' : \iota_i(w') = \iota_i(w)\}$.

Knowledge of intentions is, to say the least, a minimal requirement. Among other things, it allows agents to form intentions that they *know* are impossible to realize. Consider, for example, the epistemic frame for a Hi-Lo game depicted in Figure 5.3. At $Lo - Lo$, none of the agents consider $Hi - Hi$ possible. If we

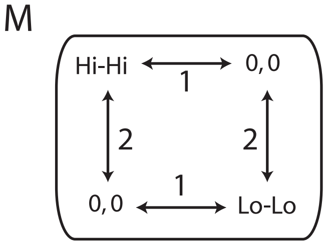

Figure 5.3: Epistemic frame for a Hi-Lo game. Only the epistemic relations are represented.

assume that one of the agents has payoff-compatible intentions, which for him boils down to intending $Hi - Hi$, then we find that this agent intends something that he himself considers impossible[13]

This is an obvious violation of belief consistency[14], according to which agents should only form intentions which are consistent with what they think the world

[13] Observe, however, that this does not preclude him from knowing what he intends. In fact, I show shortly that such cases of intention-irrationality are introspective. When the agent intends something which he considers impossible, he knows it.

[14] Recall the remarks about this in the Introduction, Section 1.2. I also revisit belief consistency in the next Chapter, Section 6.1.

is like. They should not have intentions that they think are impossible to achieve. This idea, which is stronger than knowledge of intentions, is not in itself representable in epistemic game frames for the simple reason that there is no representation of beliefs there. One can, however, look at "knowledge"-consistency of intention.

$$I_i\phi \rightarrow \Diamond_i\phi \qquad \text{(Knowledge consistency } (IK_i))$$

Put contrapositively, this formula states that agents do not form intentions to achieve facts that they know are impossible. This condition is stronger than belief consistency would be, since knowledge is always veridical in epistemic game frames. Belief consistency allows agents to form intentions that are, in fact, impossible, just as long as the achievement of these intentions is consistent with what the agent (maybe mistakenly) believes. But this cannot be the case with respect to what the agents know in epistemic game frames. Knowledge-consistency of intentions thus strongly ties intentions with what is actually true at a given state.

Knowledge consistency corresponds to the fact that, for a given state w, there is at least one w' such that $w \sim_i w'$ and $w' \in \downarrow\iota_i(w)$. In other words, there is at least one state that is compatible with the agent's intentions which he considers possible. See Figure 5.4 for an illustration of this condition.

5.2.6. Fact.

1. Take an arbitrary frame $\mathbb{F}$ in a class F. If $[w]_i \cap \downarrow\iota_i(w) \neq \emptyset$ for all w, then $\mathsf{F} \models IK_i$.

2. If for all models $\mathbb{M}$ based on a frame $\mathbb{F}$, $\mathbb{M} \models IK_i$, then $[w]_i \cap \downarrow\iota_i(w) \neq \emptyset$ for all state w in that frame.

Proof. The proof of (1) is straightforward. For (2), take a frame $\mathbb{F}$ with two states w, w' such that $\downarrow\iota_i(w) = \{w'\}$. Assume that $\mathbb{M} \models IK_i$ and that, for a given σ, we have $V(\sigma) = \{w'\}$. This means that $\mathbb{M}, w \models I_i\sigma$, and so by assumption that $\mathbb{M}, w \models \Diamond_i\sigma$, which can only be the case if $w \sim_i w'$, i.e. if $\downarrow\iota_i(w) \cap [w'] \neq \emptyset$. ■

It is worth recalling that the first account of coordination in Hi-Lo games (Section 3.5) did *not* require knowledge-consistent intentions. For this class of games, intentions can anchor coordination on the basis of a weaker constraint, namely intention-rationality. This constraint requires that, among all the profiles that can result from the strategy choice of an agent at a state, there is at least one that figures in his most precise intention (see Figure 5.5). This is also expressible in $\mathcal{L}_{\mathcal{GF}}$, as follows.

$$\bigvee_{\sigma(i)=w(i)} \mathsf{i}_i\sigma \qquad \text{(Intentions-rationality } (IR_i) \text{ at } w)$$

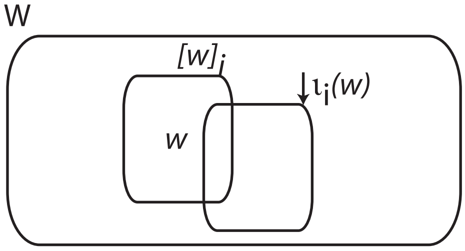

Figure 5.4: Knowledge consistency ensures that $[w]_i \cap \downarrow\iota_i(w) \neq \emptyset$.

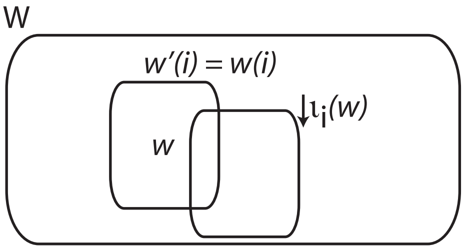

Figure 5.5: Intention-rationality ensures that $\downarrow\iota_i(w) \cap \{w' : w(i) = w'(i)\} \neq \emptyset$.

Intention-rationality and knowledge-consistency are of course related, given that agents know what they intend. First, knowledge consistency implies intention-rationality. This can easily be seen by combining Figure 5.2, 5.4 and 5.5. On the general class of epistemic game frames, however, there can be intention-rational agents who are not knowledge-consistent, as the following shows.

5.2.7. FACT. $[IK_i$ implies $IR_i]$ At any pointed model $\mathbb{M}, w$, if $\mathbb{M}, w \models IK_i$ then $\mathbb{M}, w \models IR_i$. There are, however, models where the converse does not hold.

Proof. For the first part, I show the contrapositive. Assume that $\mathbb{M}, w \models \neg IR_i$. That is, $\mathbb{M}, w \models \bigwedge_{\sigma(i)=w(i)} \neg \mathrm{i}_i\sigma$. This means that $w'(i) \neq w(i)$ for all $w' \in \downarrow\iota(w)$. But we also know, by definition, that $[w]_i \subseteq \{w' : w'(i) = w(i)\}$, which means that for all $w' \in \downarrow\iota_i(w)$, $w' \notin [w]_i$. Take s_i^* to be the collection of nominals σ' such that $V(\sigma') \in \downarrow\iota_i(w)$. We get $\mathbb{M}, w \models I_i(\bigvee_{\sigma' \in s_i^*} \sigma') \wedge K_i \neg(\bigvee_{\sigma' \in s_i^*} \sigma')$.

For the second part, take any pointed model $\mathbb{M}, w$ where $[w]_i \subset \{w' : w'(i) = w(i)\}$, and fix $\downarrow\iota_i(w) = \{w' : w'(i) = w(i)\} - [w]_i$. We obviously get that

$\mathbb{M}, w \models IR_i$. But by again using s_i^* as the collection of nominals σ' such that $V(\sigma') \in \downarrow\iota_i(w)$, we get that $\mathbb{M}, w \models I_i(\bigvee_{\sigma' \in s_i^*} \sigma') \wedge K_i\neg(\bigvee_{\sigma' \in s_i^*} \sigma')$ ■

Knowledge consistency thus implies intention rationality. This crucially rests on the fact that agents know what they choose and intend. The two notions in fact coincide when we tighten even further the connection between knowledge and strategy choices, i.e. when $w(i) = w'(i)$ if *and only if* $w \sim_i w'$, as in [van Benthem, 2003]. In other words, IK_i and IR_i are the same on game frames where the agents consider possible all the strategy profiles that can result from their strategy choices. This can easily be seen by fixing $[w]_i = \{w' : w'(i) = w(i)\}$ in Figure 5.1, and then combining it with Figure 5.4 and 5.5.

It is thus no coincidence that, in the last proof, i is knowledge-inconsistent but still intention-rational at a state where he does not consider all of his opponents' replies possible. The relation between intention rationality and knowledge consistency for an agent depends directly on what he considers possible. The following strengthening of IR_i makes this even more explicit.

$$\bigvee_{\sigma(i)=w(i)} \mathsf{i}_i\sigma \wedge \Diamond_i\sigma \qquad \text{(Epistemic intentions-rationality } (IR_i^*) \text{ at } w)$$

IR_i^* is an epistemically constrained version of IR_i. It requires not only of agents that their strategy choice somehow matches their intentions, but also that they consider the "matching" profile possible. Under knowledge of one's own strategy choice, IR_i^* is just IK_i under a different guise.

5.2.8. Fact. [Equivalence of IK_i and IR_i^*] At any pointed model $\mathbb{M}, w$, $\mathbb{M}, w \models IK_i$ iff $\mathbb{M}, w \models IR_i^*$.

Proof. Provided the first part of Fact 5.2.7, all that remains to be shown is the right-to-left direction. Again, I do it in contrapositive. Assume that $\mathbb{M}, w \models I_i\phi \wedge K_i\neg\phi$ for some ϕ. This means that $\downarrow\iota_i(w) \subseteq ||\phi||$ and that $||\phi|| \cap [w]_i = \emptyset$. We thus have $\downarrow\iota_i(w) \cap [w]_i = \emptyset$. But this means that for all σ' and $w'(i) = w(i)$ such that $V(\sigma') = w'$ and $w' \in [w]_i$, $\mathbb{M}, w \models \neg\mathsf{i}_i\sigma'$. In other words, $\mathbb{M}, w \models \bigwedge_{\sigma'(i)=s_i} \Diamond_i\sigma' \rightarrow \neg\mathsf{i}_i\sigma'$. ■

This results rests crucially on the condition that agents know their strategy choice. To see this, consider the epistemic game frame in Figure 5.6. Suppose that $w_1(1) \neq w_3(1)$, that at w_1 we have $\downarrow\iota_1(w_1) = \{w_3\}$ and that PROP is empty. We clearly get $\mathbb{M}, w_1 \models I_1\phi \rightarrow \Diamond_1\phi$ while $\mathbb{M}, w_1 \not\models IR_1$ and $\mathbb{M}, w_1 \not\models IR_1^*$.

This is a clear instance of interaction between the assumptions on epistemic game frames and intention-related conditions such as intention rationality and knowledge consistency. This interaction goes deeper, in fact. Precisely because the agents are assumed to know their own intentions and actions, both knowledge consistency and intention rationality are fully introspective. Indeed, knowledge of

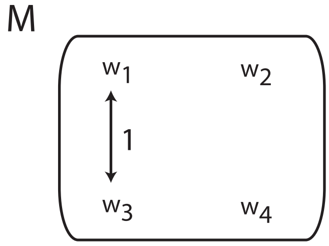

Figure 5.6: An epistemic game frame where knowledge of strategy choice is violated.

intentions and of strategy choice means that at all states that the agent considers possible he has the same intentions and chooses the same strategy. So if his strategy choice is compatible with his intentions at one of these states, it is compatible with those of his intentions at all states that he considers possible, i.e. he knows that he is intention-rational. The following makes this precise.

5.2.9. Fact. [Positive and negative introspection of IR_i and IK_i] For all pointed game models, $\mathbb{M}, w \models IR_i$ implies $\mathbb{M}, w \models K_i IR_i$ and $\mathbb{M}, w \models \neg IR_i$ implies $\mathbb{M}, w \models K_i \neg IR_i$. The same hold for IR^*_i.

Proof. I only prove positive introspection for IR_i; the arguments for the other claims are similar. Assume that $\mathbb{M}, w \models IR_i$. This happens if and only if there is a $w' \in \downarrow \iota_i(w)$ such that $w'(i) = w(i)$. Take any $w'' \in [w]_i$. We know that $\iota_i(w'') = \iota_i(w)$, which means that w' is also in $\downarrow \iota_i(w'')$, and $w''(i) = w(i)$, which means that $w''(i) = w'(i)$, and so that $\mathbb{M}, w'' \models IR_i$. ■

Introspection of intention rationality is a surprising consequence of knowledge of one's own intention. Recall that the former notion is not *per se* knowledge-related. It only refers to the connection between strategy choices and intentions. Agents are introspective about their own intention rationality because we assume that they know what they choose and what they intend.

That this assumption ensures negative introspection of both knowledge consistency and intention rationality is also surprising. With respect to knowledge and intentions, negative introspection is often seen as an overly strong condition[15]. Agents who do not intend to do something are not automatically required to know that they do not have this intention. Similarly, one might not view intention irrationality as something that agents automatically know of themselves.

[15] See e.g Wallace [2006, chap.5].

But the last fact shows that if one seeks to abandon this assumption in epistemic game frames one has to give up an apparently much weaker one, namely that agents know what they do and what they intend. In other words, with this apparently weak assumption one gives agents rather strong introspective powers.

Introspection helps to give us a better understanding of the coordination result for Hi-Lo games from Chapter 3.

5.2.10. Fact. Let *payoff-compatibility of intentions* be defined as follows:

$$\phi >^{i}_{\forall\forall} \psi \wedge I_i(\phi \vee \psi) \rightarrow \ I_i\phi \qquad \text{(Payoff-compatibility (IPC}_i\text{))}$$

Take any pointed game model $\mathbb{M}, w$ of an Hi-Lo game. Then the following are equivalent:

- $\mathbb{M}, w \models \bigwedge_{i\in I} K_i(\bigwedge_{j\neq i}(IPC_j \wedge IR_j))$
- w is the Pareto-optimal profile of that game.

Proof. The proof is essentially the same as in Chapter 3. Instead of restating the details, I only highlight what is relevant for the present analysis.

The statement of this fact uses the notion of payoff-compatibility, which in turn uses a lifted preference relation: $\phi >_{\forall\forall} \psi$. This formula states that ϕ is strictly preferred to ψ whenever all states that make ϕ true are strictly preferred to all states that make ψ true[16]. An argument very similar to the one provided for introspection on IR_i shows that payoff-compatibility of intention is also introspective.

Now, the key observation underlying the present fact is that the formula $K_i(\bigwedge_{j\neq i}(IPR_j \wedge IR_j))$ is true only when i's epistemic accessibility relation is restricted to states where all his opponents have payoff-compatible intentions and are intention-rational. Since this is *veridical* mutual knowledge, i is also intention-rational, and his intentions are payoff-compatible. Furthermore, because these two notions are introspective, i knows it. Given the structure of Hi-Lo games, this means that i's most precise intention contains exactly the Pareto-optimal profile. But since this is the case for all i, we find that the formula $K_i(\bigwedge_{j\neq i}(IPR_j \wedge IR_j))$ can only be true at that profile. ■

For the first time in this section, this result explicitly uses the preferences fragment of $\mathcal{L}_{\mathcal{GF}}$. Indeed, what is so special about the Pareto-optimal profile is its place in the preference relations: it the most preferred Nash equilibrium. That we can capture this notion in $\mathcal{L}_{\mathcal{GF}}$ should not come as a surprise, though. This language can capture Nash equilibria in the general case. We showed in van Benthem et al. [2005] how to do this using "distributed knowledge", a notion which is also definable in $\mathcal{L}_{\mathcal{GF}}$ and which "pools" the epistemic accessibility relations

[16] The precise definition of this relation can be found in the Appendix 5.5.2.

together. But $\mathcal{L}_{\mathcal{GF}}$ also allows for a characterization of this solution concept that echoes the well-known result of Aumann and Brandenburger [1995, p.1167] that I mentioned in Chapter 3 (Section 3.5).

Recall that they have shown that in two-player strategic games, if at a state each player is "weakly rational" [van Benthem, 2003, p.17] and knows his opponent's strategy choice, then at this state the agents play a Nash equilibrium. Now, the notion of "weak rationality" is not in itself expressive in $\mathcal{L}_{\mathcal{GF}}$. By introducing it as a special propositional atom[17], we can however capture in $\mathcal{L}_{\mathcal{GF}}$ the epistemic characterization of Aumann and Brandenburger [1995].

5.2.11. Definition. [Weak rationality] For a given profile $w \in W$ and strategy $s \in S_i$, take $w[s/w(i)]$ to be the profile w' that is exactly like w expect that $w'(i) = s$.

$$\mathbb{M}, w \models WR_i \quad \text{iff} \quad \begin{array}{l} \text{for all } s \in S_i \text{ such that } w(i) \neq s \\ \text{there is a } w' \sim_i w \text{ such that } w' \succeq_i w'[s/w'(i)]\ . \end{array}$$

The notion of weak rationality is easier to grasp in contrapositive. An agent is not weakly rational at a state w when one of his strategies, different from the one he plays at w, gives him strictly better outcomes in all combinations of actions of others that he considers possible. In other words, an agent is not weakly rational at a state when, as far as he knows, he plays a dominated strategy at that state.

Weak rationality thus crucially involves what the agent considers possible. He is weakly rational when he can find a "reason" to justify his current choice instead of any other options. That is, for each of his alternative strategy s' there is a state that he considers possible in which his current strategy choice is at least as good as s'. The characterization below exploits this fact by restraining what each agent considers possible. They know their own action, and so they can only be uncertain about the action of their opponent. But if they know this action too, they know the actual profile. This means that if they are weakly rational, their current strategy choice is as least as good as any other, given this strategy choice of their opponent. In other words, they are weakly rational if their strategy choice is a best response to the strategy choice of the other, which is just what Nash equilibrium requires.

5.2.12. Fact. [Nash equilibrium definability] Given a game model $\mathbb{M}$ with two agents, a profile w named by σ is a Nash equilibrium if it satisfies the following:

$$WR_1 \wedge K_1\sigma(2) \wedge WR_2 \wedge K_2\sigma(1)$$

[17] This notion could have been expressible in $\mathcal{L}_{\mathcal{GF}}$, provided I had equipped it with binders (see ten Cate [2005, p.133]). Again, I did not go in that direction in order to keep the language relatively simple.

Proof. It is enough to show that $w(i)$ is a *best response* for both agents, that is for all $s \in S_i$ and $w' = w[s/w(i)]$, $w \succeq_i w'$ for $i \in \{1, 2\}$. Consider player 1. Given $\mathbb{M}, w \models WR_1$, we will be done if we can show that $[w]_i = \{w\}$. Now observe that $\mathbb{M}, w \models K_1\sigma(2)$ is just the same as saying that for all $w' \sim_1 w$, $\mathbb{M}, w' \models \bigvee_{\sigma(2)=w(2)} \sigma$. That is, for all $w' \sim_1 w$, $w(2) = w'(2)$. But by definition we also know that $w(1) = w'(1)$ for all those w', which means that $w' = w$. The same reasoning applies to player 2, showing that w is indeed a best response for both agents. ■

This characterization of Nash equilibria, as well as that of van Benthem et al. [2005], shows that $\mathcal{L}_{\mathcal{GF}}$ can capture features of epistemic game frames not only in relation to what the agents know and intend, but also the well-known solution concepts[18]. In other words, the present framework is sufficiently encompassing to capture aspects of both instrumental and intention-based rationality, as well as the information that agents have about them in strategic interaction. What is more, this framework has a concrete deductive component, as I show now, which makes it a genuine theory of practical *reasoning* for rational planning agents.

5.2.3 Axiomatization

The set of valid formulas of $\mathcal{L}_{\mathcal{GF}}$ over the class of epistemic game frames is completely axiomatizable by the system presented in Table 5.2 on page 103. As for the axiomatization over the class of preference frames, the reader will recognize in this table many formulas that I studied in the previous section. Most of the others axiomatize the hybrid fragment of this logic. The following theorem shows that this axiom system does indeed capture reasoning in games with intentions. The details of the proof are given in Appendix 5.5.4.

5.2.13. THEOREM. *The logic $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$ is complete with respect to the class of epistemic game frames with intentions.*

5.3 Transformations of games with intentions

In the previous section I studied what can be said about what the players know, intend and ultimately decide in epistemic game frames with intentions. This covered one aspect of intention-based practical reasoning, namely how agents take their intentions and those of others into account in deliberation. But as I have already mentioned many times, this is only one part of the story. Intentions also play an active role in the shaping of decision problems.

[18] As the reader may have come to realize, most of this high expressive power comes from the "hybrid" fragment, i.e. the constants that directly name strategy profiles. I show in Appendix 5.5.3 that nominals are indeed crucial to capture Nash equilibria.

- All propositional tautologies.
- $S4$ for $\Diamond^{\leq}$ and, for all σ and σ':
  (Tot) $(\sigma \wedge \Diamond_i^{\leq}\sigma') \vee (\sigma' \wedge \Diamond_i^{\leq}\sigma)$
- For $\Diamond^{<}$, K and:
  (Irr) $\sigma \rightarrow \neg\Diamond^{<}\sigma$
  (Trans) $\Diamond^{<}\Diamond^{<}\sigma \rightarrow \Diamond^{<}\sigma$
- For I_i:
  (K) $I_i(\phi \wedge \psi) \leftrightarrow I_i\phi \wedge I_i\psi$
  (Ser) $\mathsf{i}_i\top$
- $S5$ for E.
- Interaction axioms.
  (Exists$_\sigma$) $<> \phi \rightarrow E\phi$
  (Inc$_{E-\sigma}$) $E(\sigma \wedge \phi) \rightarrow A(\sigma \rightarrow \phi)$
  (Inc$_\sigma$) $E(\sigma)$
  (Inc$_1$) As in Section 5.1.3.
  (K-I) $I_i\phi \rightarrow K_iI_i\phi$
- (Nec) for all modal connective, and the following additional inference rules. In both cases $\sigma \neq \sigma'$ and the former does not occur in ϕ.
  - (Name) From $\sigma \rightarrow \phi$ infer ϕ.
  - (Paste) From $(E(\sigma' \wedge <> \sigma) \wedge E(\sigma \wedge \phi)) \rightarrow \psi$ infer $E(\sigma' \wedge <> \phi) \rightarrow \psi$

Table 5.2: The axiom system for $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$. Here $<>_i$ is any of $\Diamond_i$, $\Diamond_i^{<}$, $\Diamond_i^{\leq}$ or i_i.

In Chapter 4 I modelled this process with two transformations of strategic games: *cleaning*, which excluded options that are inconsistent with one's intentions, and *pruning*, in which irrelevant details were overlooked. In this section I use logical methods to gain further insights into cleaning. I show that altruistic cleaning naturally relates to the notion of intention rationality that I used in the previous section. This observation opens the door to a whole family of cleaning-like operations, definable using a *dynamic* extension of $\mathcal{L}_{\mathcal{GF}}$, while at the same time giving a more general perspective on transformations of epistemic game frames.

5.3.1 Dynamic language

I introduced cleaning as, so to speak, a dynamic add-on that supplements the machinery of strategic games with intentions. That is, the exclusion of intention-inconsistent options comes, as a separate module, on top of the "static" analysis of games in terms of knowledge, preferences and intentions. In the same modular manner, *dynamic* epistemic logic (DEL) extends "static" epistemic languages to talk about information changes[19]. The DEL approach is thus a natural environment for broadening our perspective on cleaning-like operations in epistemic game frames.

In its full generality, DEL can analyze the most diverse information changes. However, operations like cleaning, which contract relational frames, correspond to a very simple fragment of DEL, known as *public announcements logic*. In this logic the contraction of a relational model is viewed as the result of publicly and truthfully announcing that a given formula is true. More precisely, in a given relational model, announcing that ϕ means looking only at the sub-model where ϕ holds. A public announcement formula, denoted $[\phi!]\psi$, thus says that after removing from the original models all the states where ϕ does not hold, ψ is the case. As we shall see shortly, cleaning will indeed correspond to the announcement of a particular formula of $\mathcal{L}_{\mathcal{GF}}$. But before showing this, let me look at the "public announcement" extension of $\mathcal{L}_{\mathcal{GF}}$ in full generality.

5.3.1. Definition. [Dynamic extension of $\mathcal{L}_{\mathcal{GF}}$] $D\mathcal{L}_{\mathcal{GF}}$, the dynamic extension of $\mathcal{L}_{\mathcal{GF}}$, is defined as follows :

$$\phi ::= p \mid \sigma \mid \phi \wedge \phi \mid \neg\phi \mid \Diamond^{\leq}\phi \mid \Diamond^{<}\phi \mid K_i\phi \mid I_i\phi \mid \; E\phi \mid [\phi!]\phi$$

The only new formulas in this language are of the form $[\phi!]\psi$. They should be read as "after truthfully announcing that ϕ, it is the case that ψ". As I have just written, these announcements correspond to contractions of the underlying epistemic game model, that I denote $\mathbb{M}_{|\phi}$.

5.3.2. Definition. [Contracted epistemic game models] Given an epistemic game model $\mathbb{M}$ and a formula $\phi \in D\mathcal{L}_{\mathcal{GF}}$, the *contracted* model $\mathbb{M}_{|\phi}$ is defined as follows[20].

1. $W_{|\phi} = ||\phi||$.
2. $\sim_{i\,|\phi}$ is the restriction of $\sim_i$ to $W_{|\phi}$. Similarly for $\succeq_i$.
3. $\iota_{|\phi}(w) = \begin{cases} \uparrow (||\phi|| \cap \downarrow\!\iota(w)) & \text{if } ||\phi|| \cap \downarrow\!\iota(w) \neq \emptyset \\ W_{|\phi} & \text{otherwise} \end{cases}$

[19] Key references on the topic are Plaza [1989], Gerbrandy [1999], Baltag et al. [1998] and van Ditmarsch et al. [2007].

[20] In this definition $\uparrow\! A$, for a given set $A \subseteq W$, is a shorthand for the closure under supersets of A, that is $\uparrow\! A = \{B : A \subseteq B \subseteq W\}$. In item (3) the closure is with respect to $W_{|\phi}$.

4. $V_{|\phi}$ is the restriction of V to $W_{|\phi}$.

The domain $W_{|\phi}$ of a model restricted to ϕ is just what one would expect. It is the set of states where ϕ was the case before the announcement. The epistemic and preferences relations are modified accordingly.

The restriction of the intention function ι_i splits into two cases. On the one hand, if what is announced was compatible with the agent's intention, that is if $||\phi|| \cap \downarrow\iota(w) \neq \emptyset$, then the agent "restricts" his intention according to the announcement, just as the agents restricted their intentions after cleaning or pruning in Chapter 4. Formally, the new intention set is built just as before, by taking restriction of the most precise intention to the states compatible with the formula announced : $\iota_{|\phi}(w) = \uparrow(||\phi|| \cap \downarrow\iota(w))$. This process is illustrated in Figure 5.7. For the other case, where the announcement is *not* compatible with what the agent intends, that is when $\downarrow\iota_i(w) \cap ||\phi|| = \emptyset$, I introduce here an elementary intention revision policy. The agent conservatively bites the bullet, so to speak. He indeed throws away the old, unachievable intentions but, on the other hand, he refrains from committing to anything other than what he already knows to be the case. In other words, the agent's intention revision boils down to his not forming any new specific intentions, which formally gives $\iota_{|\phi}(w) = \{W_{|\phi}\}$. This is illustrated in Figure 5.8.

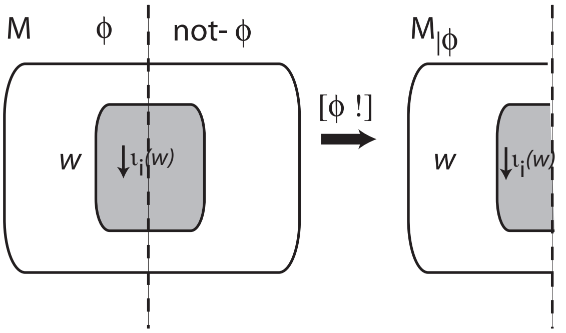

Figure 5.7: The intention restriction when $\downarrow\iota_i(w) \cap ||\phi|| \neq \emptyset$.

I do not claim that this revision policy is "right" or adequate. It is a simple starting point, using only existing resources of epistemic game frames. In Section 5.3.3 I shall have many occasions to observe how this policy behaves, and will then be in a better position to assess it.

A model $\mathbb{M}_{|\phi}$ restricted after the announcement of ϕ is thus built out of the sub-model of $\mathbb{M}$ where ϕ holds before the announcement, with the valuation, the epistemic accessibility relations, the preferences and the intention functions restricted accordingly. Equipped with such models, we can give a generic definition of the truth condition of formulas of the form $[\phi!]\psi$.

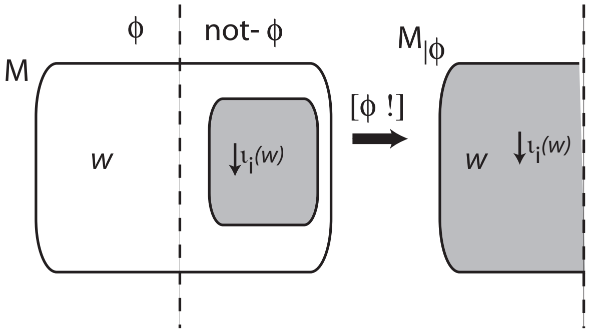

Figure 5.8: The intention restriction when $\downarrow\iota_i(w) \cap ||\phi|| = \emptyset$.

5.3.3. Definition. [Truth for public announcement formulas]

$$\mathbb{M}, w \models [\phi!]\psi \quad \text{iff} \quad \text{If } \mathbb{M}, w \models \phi \text{ then } \mathbb{M}_{|\phi}, w \models \psi.$$

The condition "If $\mathbb{M}, w \models \phi$ then... " ensures that we are dealing with *truthful* announcements. That is, only true facts can be announced publicly in this logic. This simplification will not hamper the present analysis. Cleaning was always based on actual, i.e. veridical, intentions of agents, as will the other cleaning-like operations that $D\mathcal{L}_{\mathcal{GF}}$ unveils.

It is important to observe at this point that announcements are not necessarily "successful"—I give a precise definition of this notion later on. It can happen that an announcement is self-refuting, in the sense that announcing it truthfully as a formula makes it false. This is so because even though announcing true things does not change the "basic" facts of the situation, it definitely changes the *information* that agents have about these facts[21]. When an announcement contains informational facts, it can thus make these very facts false.

As in the last two sections, these announcements are studied by looking at the expressive power of $D\mathcal{L}_{\mathcal{GF}}$. But, unlike what I did in these sections, I first look at the logic of public announcements in epistemic game frames. The axiomatization techniques for this logic are slightly different from what we have seen so far, and also provide tools for what comes thereafter.

5.3.2 Axiomatization

The logics $\Lambda_{\mathcal{L}_{\mathcal{P}}}$ and $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$ were devised in a very similar and also quite standard manner. They consisted of a set of axioms encapsulating properties of the intended class of frames, for example $\Diamond^{\leq}\Diamond^{\leq}\phi \rightarrow \Diamond^{\leq}\phi$ for transitivity of $\succeq$, together

[21] See van Benthem [2006a] for more on this phenomenon.

with some inference rules. The completeness arguments for these logics were also quite standard.

The axiomatization of the valid formulas of $D\mathcal{L}_{\mathcal{GF}}$ over the class of epistemic game frames proceeds differently. One does not need directly to provide formulas that correspond to properties of public announcements[22]. Rather, it is enough to provide a set of formulas, shown in Table 5.3, that allow one to compositionally translate formulas with $[\phi!]$ operators to formulas in $\mathcal{L}_{\mathcal{GF}}$. If one can show that these formulas are valid, completeness of $\Lambda_{D\mathcal{L}_{\mathcal{GF}}}$ with respect to the class of epistemic game frames is just a corollary of completeness of $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$ with respect to this very class of frames. By taking these formulas as axioms of $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$, valid formulas in $D\mathcal{L}_{\mathcal{GF}}$ are then deductively reducible to valid formulas of $\mathcal{L}_{\mathcal{GF}}$, which we know can in turn be deduced in $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$. In other words, the axioms of Table 5.3 show that agents can reason about information change in games with intentions in the basis of what information they have about each other's knowledge and intentions.

The detailed arguments for the validity of the formulas in Table 5.3 can be found in Appendix 5.5.5. They all explain post-announcement conditions in terms of pre-announcement ones. Let me look briefly at (5), which encodes the intention revision policy that I have just discussed. It states the pre-announcement conditions under which it can be the case that, after an announcement that ϕ, i intends that ψ. Not surprisingly, these conditions match the two cases of the update rule for ι_i. If the intentions of i were already compatible with the announcements, that is if $\mathsf{i}_i\phi$, then one should have been able to find $||\phi||$ in the intention set of i, once restricted to $||\phi||$. This is essentially what $I_i(\phi \rightarrow [\phi!]\psi)$ says. On the other hand, if the announcement of ϕ was not compatible with ϕ, i.e. if $\neg\mathsf{i}_i\phi$, then i intends ψ after the announcement if and only if ψ is true everywhere in the restricted model, i.e. $[\phi!]A\psi$, which is exactly what the intention revision rule for ι_i prescribes.

1. $[\phi!]x \leftrightarrow \phi \rightarrow x$ for $x \in \textsc{prop} \cup S$.
2. $[\phi!]\neg\psi \leftrightarrow \phi \rightarrow \neg[\phi!]\psi$.
3. $[\phi!]\psi \wedge \xi \leftrightarrow \phi \rightarrow ([\phi!]\psi \wedge [\phi!]\xi)$.
4. $[\phi!][\cdot]\psi \leftrightarrow \phi \rightarrow [\cdot](\phi \rightarrow [\phi!]\psi)$
5. $[\phi!]I_i\psi \leftrightarrow \phi \rightarrow (\mathsf{i}_i\phi \wedge I_i(\phi \rightarrow [\phi!]\psi) \vee (\neg\mathsf{i}_i\phi \wedge [\phi!]A\psi))$

Table 5.3: The axiom system for $\Lambda_{D\mathcal{L}_{\mathcal{GF}}}$. Here $[\cdot]$ is either A, K_i, $\Box_i^{\leq}$ or $\Box_i^{<}$.

[22] Although the following set of axioms in Table 5.3 can also be seen as tightly fixing the properties of model restrictions. See the references in the footnote on page 13, in the Introduction.

As we shall see in the next section, these axioms not only provide a "shortcut" toward completeness results; they also prove very useful in understanding pre-announcement conditions.

5.3.3 Expressive power

The dynamic extension of $\mathcal{L}_{\mathcal{GF}}$ really unfolds the full expressive power of this language. It provides a unified framework for studying practical reasoning of planning agents in strategic interaction. In short, it gets us closer to the "big picture" of intention-based practical reasoning.

This section is quite long, in comparison to the preceding ones. I have thus split it into three parts. First, I look back at altruistic cleaning from the perspective of $D\mathcal{L}_{\mathcal{GF}}$. Second, I investigate more closely the behaviour of the intention revision policy in the context of overlapping intentions. Finally, I look at enabling conditions for cleaning, in terms of announcements of weak rationality. Of course, much more could be said using the dynamic language about how information changes in games with intentions. I have chosen these three topics because, on the one hand, they shed new light on phenomena that we have encountered previously and, on the other hand, because they are paradigmatic of the type of analysis that can be conducted with dynamic epistemic logic on epistemic game frames.

Varieties of cleaning

Let me start by looking at the cleaning of decision problems. Both versions of this operation, individualistic and altruistic, were defined with respect to a given intention profile, one for the whole decision problem. By contrast, in epistemic game frames we have "state-dependent" intentions, that is one intention profile per state. There is a further difference between the decision problem I used in Chapter 4 and the current epistemic game frames. As mentioned at the beginning of Section 5.2.1, I distinguished strategy profiles and outcomes in the former, but not in the latter. To capture cleaning in $D\mathcal{L}_{\mathcal{GF}}$, I have to take care of these two differences.

The second one requires a slight redefinition of the cleaning operation to fit epistemic game frames. For most of this section I shall focus on altruistic cleaning. As we shall see later, one can present a very similar analysis of individualistic cleaning.

5.3.4. Definition. [Altruistic cleaning of epistemic game frame] Given an epistemic game frame $\mathbb{G}$ and an intention profile ι, the *cleaned strategy set* $cl(S_i)$ for an agent i is defined as

$$cl(S_i) = \{s_i \mid \text{there is a } w' \in \downarrow\iota_i \text{ such that } w'(i) = s_i\}$$

The *altruistically cleaned* version of $\mathbb{G}$ from the point of view of the intention profile ι, denoted $cl_\iota(\mathbb{G})$, is defined as follows.

- $cl(W) = \{w \mid \exists i \text{ such that } w(i) \in cl(S_i)\}$.
- $\sim_i^{cl}$, $\succeq_i^{cl}$ and V^{cl} are restriction of $\sim_i$, $\succeq_i$ and V^{cl} to $cl(W)$, respectively.
- For all i, $\iota_i^{cl} = \uparrow(cl(W) \cap \downarrow\iota_i)$.

The second point of divergence between the strategic games of Chapter 4 and epistemic game frames is also easily accommodated.

5.3.5. Definition. [State-independent intentions] An epistemic game frame $\mathbb{G}$ is said to have *state-independent* intentions whenever, for all $w, w' \in W$ and $i \in I$, $\iota_i(w) = \iota_i(w')$.

Altruistic cleaning can thus be seen as an operation on epistemic game frames with state-dependent intentions. With this in hand, we can readily characterize cleaning in $D\mathcal{L}_{\mathcal{GF}}$. It corresponds to the public announcement of a crucial concept: intention-rationality.

5.3.6. Fact. For any model $\mathbb{M}$ with state-independent intentions, its cleaned version $cl_\iota(DP)$ is exactly the model that results from announcing $\bigvee_{i\in I} IR_i$.

Proof. I first show that $cl(W) = W_{|\bigvee_{i\in I} IR_i}$. We know that $w' \in cl(W)$ iff there is an i such that $w'(i) \in cl(S_i)$. This, in turn, happens iff there is an i and a $w'' \in \downarrow\iota_i$ such that $w''(i) = w'(i)$, which is also the same as to say that there is an i such that $\mathbb{M}, w' \models \bigvee_{\sigma(i)=w'(i)} \mathsf{i}_i\sigma$. This last condition is equivalent to $\mathbb{M}, w' \models \bigvee_{i\in I} IR_i$, which finally boils down to $w' \in W_{|\bigvee_{i\in I} IR_i}$. It should then be clear that the restricted cleaned relations and valuation correspond to those obtained from the announcement of $\bigvee_{i\in I} IR_i$, and vice versa. It remains to be shown that the two operations update the intention sets similarly. Here the state-independence becomes crucial, witness the following:

5.3.7. Lemma. *For any state-independent intention ι_i:*

$$\downarrow\iota_i \cap \left\|\bigvee_{i\in I} IR_i\right\| \neq \emptyset$$

Proof. Take any such ι_i. We know that $\downarrow\iota_i \neq \emptyset$. So take any $w \in\downarrow\iota_i$. We have that $\mathbb{M}, w \models \mathsf{i}_i\sigma$ for $V(\sigma) = w$. But since we are working with state independent intentions, $\mathsf{i}_i\sigma$ implies $\bigvee_i IR_i$, as can be seen by unpacking the definition of the latter. This means that $w \in || \bigvee_{i\in I} IR_i||$ and thus that $\downarrow\iota_i \cap || \bigvee_{i\in I} IR_i|| \neq \emptyset$. ■

This lemma reveals that for state-independent intentions, the second clause of the definition of $\iota_{i|\phi}$ is never used. But then it should be clear that for all i, $\iota_i^{cl} = \iota_{i|\bigvee_{i\in I} IR_i}$.

■

This characterization of cleaning in terms of intention-rationality shows that the two notions are really two sides of the same coin: altruistically inadmissible options are just intention-irrational ones, and vice versa. This characterization also highlights the *altruistic* aspect of cleaning. That the operation corresponds to the announcement of a disjunction over the set of agents is indeed quite telling. The idea behind altruistic cleaning was that the agents retained all the strategies which were compatible with the intentions of *one* of their co-players. This is exactly what the announcement says: *one* of the agents is intention-rational. Along the same line, an easy check reveals that cleaning with individualistic admissibility can be characterized in terms of a stronger, i.e. *conjunctive* announcement of intention-rationality.

Recall that intention-rationality is introspective (Fact 5.2.9). Agents in epistemic game frames know whether they are intention-rational at a given state. The following shows that in epistemic game frames with state-independent intention, announcing intention-rationality is, so to speak, safe. Intention-rationality is robust to altruistic cleaning[23].

5.3.8. Definition. [Self-fulfilling announcements] An announcement that ϕ is said to be *self-fulfilling* at a pointed model $\mathbb{M}, w$ if $\mathbb{M}, w \models [\phi!]A\phi$.

5.3.9. Fact. The announcement of $\bigvee_{i \in I} IR_i$ is self-fulfilling for any pointed model $\mathbb{M}, w$ with state-independent intentions.

Proof. We have to show that for any pointed model with state-independent intentions, if $\mathbb{M}, w \models \bigvee_{i \in I} IR_i$ then $\mathbb{M}_{|\bigvee_{i \in I} IR_i}, w \models A \bigvee_{i \in I} IR_i$. I will show something stronger, namely that for all $w \in ||IR_i||$, $\mathbb{M}_{|IR_i}, w \models A\ IR_i$.

Take any $w' \in W_{|IR_i}$. We have to show that there is a w'' in $\downarrow\iota_{i|IR_i}(w')$ such that $w''(i) = w'(i)$. Because $w' \in W_{|IR_i}$ we know that w' was in $||IR_i||$ before the announcement. But this means that there was a $w'' \in \downarrow\iota_i(w')$ such that $w''(i) = w'(i)$. But since we have state-independent intentions, this means that w'' was also in $||IR_i||$. Furthermore, that means that $\downarrow\iota_{i|IR_i}(w') = \downarrow\iota_{i|IR_i}(w') \cap ||IR_i||$, and so that $w'' \in \downarrow\iota_{i|IR_i}(w')$, as required. ■

In the course of this proof I have shown that the formula $IR_i \rightarrow [IR_i!]A(IR_i)$ is *valid* on the class of epistemic game frames with state-independent intentions. It states that if an agent is intention-rational then he remains so after the announcement of this fact. Here we can draw some interesting conclusions about intentions-based, practical *reasoning*, given the completeness result mentioned in Section 5.3.2. This formula is *not* an axiom of this logic, but by completeness we know that it is a *theorem*, that is, it is part of the information that planning agents can *deduce* from state-independence and knowledge of intentions and actions.

[23] Recall that this is not generally the case for announcements. This result is clearly the epistemic counterpart of the fixed-point result for altruistic cleaning (Fact 4.1.9).

Fact 5.3.9 thus shows that ruling out options is an operation on decision problems that agents can safely perform. Colloquially, the previous considerations show that agents who exclude inconsistent options from a given decision problem know what they are doing, and cannot mess things up by doing it. This characterization of cleaning in terms of announcement also opens the door to new kinds of cleaning operations. One can refine the notion of admissibility of options by playing with the announced formula. An obvious candidate for such a refined cleaning operation is the announcement of knowledge-consistency of intentions. Being also introspective, this notion is something that agents can knowingly announce. For the same reason, it is also "safe". If an agent has knowledge-consistent intentions at a state, he also has knowledge-consistent intentions at all states which he considers possible. But then the announcement of knowledge-consistency keeps all these states—and by the same token keeps him—knowledge-consistent. In other words, announcing knowledge consistency cannot be self-defeating. It is also robust to its corresponding operation.

5.3.10. Fact. The announcement of $\bigvee_{i \in I} IR_i^*$ is self-fulfilling for any pointed model $\mathbb{M}, w$ with state-independent intentions.

Proof. The proof follows the same line as in the previous fact. Namely, I show that if $\mathbb{M}, w \models IR_i^*$ then $\mathbb{M}_{|\bigvee_{i \in I} IR_i^*}, w \models A\ IR_i^*$. The reasoning is entirely similar. Take any $w' \in W_{|IR^*}$. We know that $\downarrow\iota_i(w') \cap [w']_i \neq \emptyset$. Now take any w'' in this intersection. Because we work with state independent intentions, we know that $\iota_i(w'') = \iota_i(w')$ and because $w'' \sim_i w'$ we know that $w''(i) = w'(i)$. Furthermore, because $\sim_i$ is an equivalence relation we know that $[w'']_i = [w']_i$. This means that $w'' \in ||IR_i^*||$. This gives us that $\downarrow\iota_{i|IR_i^*}(w') = \iota_i(w') \cap ||IR_i^*||$ and also that $w'' \in\downarrow\iota_{i|IR_i^*}(w') \cap [w']_{i|IR_i^*}$, as required. ■

In proving this fact I also show that agents remain knowledge-consistent after the announcement of this fact. Once again, it is worth stressing that planning agents in strategic games with intentions can *deduce* this. In other words, the proof of this fact unveils another valid formula to which corresponds explicit reasoning which agent performs in games with intentions.

Let me call *epistemic cleaning* the operation that corresponds to the announcement of knowledge-consistency. As one can expect from Section 5.2.2, there is a tight connection between epistemic and non-epistemic, that is intention-rationality-based cleaning. All the profiles that survive the first operation would survive the altruistic cleaning. Moreover, no further transformation can be achieved by altruistic cleaning after an epistemic one.

5.3.11. Fact. For any pointed model $\mathbb{M}, w$ with state-independent intentions:

$$\Big\| \bigvee_{i \in I} IR_i^* \Big\|_{|\bigvee_{i \in I} IR_i^*} \subseteq \Big\| \bigvee_{i \in I} IR_i \Big\|_{|\bigvee_{i \in I} IR_i}$$

Furthermore, there exist pointed models where this inclusion is strict.

Proof. The first part follows directly from Lemma 5.3.9, 5.3.10 and 5.2.7. For the second part, take the model in Figure 5.9, with $\iota_1 = \{w_1, w_3\}$ and $\iota_2 = \{w_1, w_4\}$. Observe that no state is ruled out by altruistic cleaning. But w_2 is eliminated by epistemic cleaning. Indeed, we have $\mathbb{M}, w_2 \models \neg IR_1^* \wedge \neg IR_2^*$. ■

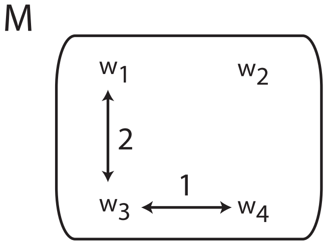

Figure 5.9: The game for the proof of Fact 5.3.11. Only the epistemic relations are represented.

For epistemic game frames with state-independent intentions, the "original" environment of cleaning, the static connection between intention-rationality and knowledge-consistency thus carries through to their dynamic counterparts. But what about the more general case of state-dependent intentions? In this more general framework the two types of cleaning are distinguished more sharply. Only knowledge consistency remains robust to its corresponding announcement.

5.3.12. Fact. The announcement of $\bigvee_{i \in I} IR_i^*$ is self-fulfilling for any pointed model $\mathbb{M}, w$.

Proof. Inspecting the proof of Lemma 5.3.10 reveals that, in fact, I did not need to use the state-independence of intention to conclude that $\iota_i(w'') = \iota_i(w')$. This was already ensured by the fact that $w'' \sim_i w'$. ■

5.3.13. Fact. The announcement of $\bigvee_{i \in I} IR_i$ is not self-fulfilling for arbitrary pointed model $\mathbb{M}, w$.

Proof. Take again the set of states in Figure 5.9, but fix the intentions as in Table 5.4. The announcement of $\bigvee_{i \in I} IR_i$ removes w_2 and w_3, making both agents intention-irrational at w_1. ■

This shows that non-epistemic cleaning is more sensitive to state-dependent intentions than its epistemic variant. Again, in more colloquial terms, one announcement of intention-rationality can mess things up when agents have state-dependent intentions. But, interestingly enough, this is not the case *in the long*

w	$\downarrow\iota_1(w)$	$\downarrow\iota_2(w)$
w_1	w_2, w_4	w_3, w_4
w_2	w_3	w_1
w_3	w_2	w_4
w_4	w_4	w_4

Table 5.4: The state-dependent intentions for Figure 5.9.

run. That is, announcing intention-rationality is self-fulfilling if it is repeated often enough, so to speak. To see this requires a few preparatory facts.

As the remark in Section 5.3.1 already suggested, I introduced the intention-revision policy precisely to avoid cases where a truthful announcement would leave the agent with inconsistent intentions. This is in fact something that agents can explicitly deduce in epistemic game frames.

5.3.14. Fact. $\mathbb{M} \models \bigwedge_{i \in I} [\phi!]\mathsf{i}_i\top$ for all models for game structure $\mathbb{M}$.

Proof. This could be shown semantically, going through the various clauses of the definition of cleaned models. Here, however, I can put the axioms from Section 5.3.2 to work to show that $\bigwedge_{i \in I}[\phi!]\mathsf{i}_i\top$ is valid. In this proof the numbers refer to Table 5.3 on page 107.

We start with $[\phi!]\neg I_i\bot$, which is the same as $[\phi!]\mathsf{i}_i\top$. By (2), this is equivalent to :

$$\phi \to \neg[\phi!]I_i\bot$$

Now, by (5), the consequent expends into two parts $\Phi = \mathsf{i}_i\phi \wedge I_i(\phi \to [\phi!]\bot)$ and $\Psi = \neg\mathsf{i}_i\phi \wedge [\phi!]A\bot$, that I treat separately to keep the formulas readable.

$$\phi \to \neg(\phi \to (\Phi \vee \Psi))$$

Before looking at each disjunct, some redundancy can be eliminated by propositional reasoning, to get:

$$\phi \to \neg(\Phi \vee \Psi)$$

Now let us first look at $\Phi = \mathsf{i}_i\phi \wedge I_i(\phi \to [\phi!]\bot)$. By (1)—$\bot$ can be treated as a propositional atom—we get:

$$\mathsf{i}_i\phi \wedge I_i(\phi \to (\phi \to \bot))$$

This is equivalent in propositional logic to:

$$\mathsf{i}_i\phi \wedge I_i(\neg\phi)$$

But the second conjunct is just the negation of the first, which means that Φ is just equivalent to $\bot$. We are thus left with :

$$\phi \to \neg(\bot \vee \Psi)$$

Which is just the same as :

$$\phi \to \neg\Psi$$

Now, recall that B is the following:

$$\neg \mathsf{i}_i\phi \wedge [\phi!]A\bot$$

By (4), this expands to:

$$\neg \mathsf{i}_i\phi \wedge A(\phi \to [\phi!]\bot)$$

By (1) again, we thus get:

$$\neg \mathsf{i}_i\phi \wedge A(\phi \to (\phi \to \bot))$$

This again reduces to:

$$\neg \mathsf{i}_i\phi \wedge A(\neg\phi)$$

Putting this back in the main formula, we get:

$$\phi \to \neg(\neg \mathsf{i}_i\phi \wedge A(\neg\phi))$$

But then propositional reasoning gets us:

$$(\phi \wedge A\neg\phi) \to \mathsf{i}_i\phi$$

But the antecedent is just a contradiction of the axiom (Ref) for E, and so we get:

$$\bot \to \mathsf{i}_i\phi$$

Which is a tautology. Since we took an arbitrary i, we can conclude that $\bigwedge_{i\in I}[\phi!]\mathsf{i}_i\top$ is also one. ■

As just stated, this result tells us that the intention revision policy that I introduce does indeed preserve the consistency of plans, and that planning agents can deduce this in epistemic game frames. But it also bears important consequences for the fixed-point behaviour of non-epistemic cleaning.

5.3.15. Definition. [Announcement stabilization] Given a pointed game model $\mathbb{M}, w$, let $\mathbb{M}^k_{|\phi}, w$ be the pointed model that results after announcing k times ϕ at w. The announcement of ϕ *stabilizes* at k for $\mathbb{M}, w$ whenever $\mathbb{M}^k_{|\phi}, w = \mathbb{M}^{k+1}_{|\phi}, w$.

To show that non-epistemic cleaning is self-fulfilling at the fixed point, I first have to show that it indeed reaches such a point.

5.3.16. Fact. [Stabilization of $[\bigvee_{i\in I} IR_i!]$] For any pointed model $\mathbb{M}, w$, the announcement of $\bigvee_{i\in I} IR_i$ stabilizes at some k.

Proof. Assume that there is no such k.[24] This means that there is no k such that $\mathbb{M}^k_{|\bigvee_{i\in I} IR_i}, w = \mathbb{M}^{k+1}_{|\bigvee_{i\in I} IR_i}, w$. Since we work with finite models, this means that there is a finite n-step loop where $\mathbb{M}^k_{|\bigvee_{i\in I} IR_i} = \mathbb{M}^{k+n+1}_{|\bigvee_{i\in I} IR_i}$ such that

$$\mathbb{M}^k_{|\bigvee_{i\in I} IR_i}, w \neq \mathbb{M}^{k+1}_{|\bigvee_{i\in I} IR_i}, w \neq ... \neq \mathbb{M}^{k+n}_{|\bigvee_{i\in I} IR_i} \neq \mathbb{M}^{k+n+1}_{|\bigvee_{i\in I} IR_i}$$

Now, observe that by Definition 5.3.2:

$$W^k_{|\bigvee_{i\in I} IR_i} \supseteq W^{k+1}_{|\bigvee_{i\in I} IR_i} \supseteq ... \supseteq W^{k+n}_{|\bigvee_{i\in I} IR_i} \supseteq W^{k+n+1}_{|\bigvee_{i\in I} IR_i}$$

But since $\mathbb{M}^k_{|\bigvee_{i\in I} IR_i} = \mathbb{M}^{k+n+1}_{|\bigvee_{i\in I} IR_i}$, all these inclusion are in fact equalities.

$$W^k_{|\bigvee_{i\in I} IR_i} = W^{k+1}_{|\bigvee_{i\in I} IR_i} = ... = W^{k+n}_{|\bigvee_{i\in I} IR_i} = W^{k+n+1}_{|\bigvee_{i\in I} IR_i}$$

Given the definition of the relations $\sim_i$ and $\succeq_i$, it must then be that for all $0 \leq \ell \leq n$, there is a $i \in I$ and a $w \in W^{k+\ell}_{|\bigvee_{i\in I} IR_i}$ such that $\iota^{k+\ell}_{i,|\bigvee_{i\in I} IR_i}(w) \neq \iota^{k+\ell+1}_{i\,|\bigvee_{i\in I} IR_i}(w)$. But this cannot be, as the following two cases show, and so there cannot be such a loop.

1. Assume that:

$$\downarrow\iota^{k+\ell}_{i|\bigvee_{i\in I} IR_i}(w) \cap \left\|\bigvee_{i\in I} IR_i\right\|^{k+\ell} \neq \emptyset \tag{1}$$

   This means that:

$$\iota^{k+\ell+1}_{i\,|\bigvee_{i\in I} IR_i}(w) = \uparrow\left(\downarrow\iota^{k+\ell}_{i|\bigvee_{i\in I} IR_i}(w) \cap \left\|\bigvee_{i\in I} IR_i\right\|^{k+\ell}\right)$$

   But observe that, while $W^{k+\ell+1}_{|\bigvee_{i\in I} IR_i} = W^{k+\ell}_{|\bigvee_{i\in I} IR_i}$:

$$\left\|\bigvee_{i\in I} IR_i\right\|^{k+\ell} = W^{k+\ell+1}_{|\bigvee_{i\in I} IR_i}$$

   This means that:

$$\iota^{k+\ell+1}_{i|\bigvee_{i\in I} IR_i}(w) = \uparrow(\downarrow\iota^{k+\ell}_{i|\bigvee_{i\in I} IR_i}(w) \cap W^{k+\ell}_{|\bigvee_{i\in I} IR_i}) = \uparrow\downarrow\iota^{k+\ell}_{i|\bigvee_{i\in I} IR_i}(w) = \iota^{k+\ell}_{i|\bigvee_{i\in I} IR_i}(w)$$

   So (1) cannot hold while:

$$\iota^{k+\ell}_{i|\bigvee_{i\in I} IR_i}(w) \neq \iota^{k+\ell+1}_{i|\bigvee_{i\in I} IR_i}(w)$$

[24] If we could show that this announcement is a *monotone map*, i.e. if it were the case that $\mathbb{M}_{|\bigvee_{i\in I} IR_i} \subseteq \mathbb{M}'_{|\bigvee_{i\in I} IR_i}$ provided that $\mathbb{M} \subseteq \mathbb{M}'$, then we would be done. See Apt [2007]. Unfortunately this announcement is not monotonic. We indeed have that $W_{|\bigvee_{i\in I} IR_i} \subseteq W'_{|\bigvee_{i\in I} IR_i}$ if $\mathbb{M} \subseteq \mathbb{M}'$. The non-monotonicity lies in the update rule for the intention set. It is not very complicated to devise an example where $\mathbb{M} \subseteq \mathbb{M}'$ but in which there is a $w \in W$ and an $i \in I$ such that $\iota_{i|\bigvee_{i\in I} IR_i}(w) \not\subseteq \iota'_{i|\bigvee_{i\in I} IR_i}(w)$. For this reason I show the existence of the fixed point directly.

2. Assume then that:

$$\downarrow\iota^{k+\ell}_{i|\bigvee_{i\in I}IR_i}(w)\cap\left\|\bigvee_{i\in I}IR_i\right\|^{k+\ell}=\emptyset \tag{2}$$

In this case $\iota^{k+\ell+1}_{i|\bigvee_{i\in I}IR_i}(w)$ just becomes $\{W^{k+\ell+1}_{|\bigvee_{i\in I}IR_i}\}$. But recall that by definition, $W^{k+\ell+1}_{|\bigvee_{i\in I}IR_i}$ is just $||\bigvee_{i\in I}IR_i||^{k+\ell}$. But since we know that $W^{k+\ell+1}_{|\bigvee_{i\in I}IR_i}=W^{k+\ell}_{|\bigvee_{i\in I}IR_i}$, this means that $||\bigvee_{i\in I}IR_i||^{k+\ell}=W^{k+\ell}_{|\bigvee_{i\in I}IR_i}$. But that would mean:

$$\downarrow\iota^{k+\ell}_{i|\bigvee_{i\in I}IR_i}(w)\cap W^{k+\ell}_{|\bigvee_{i\in I}IR_i}=\emptyset$$

Which is just to say that

$$\downarrow\iota^{k+\ell}_{i|\bigvee_{i\in I}IR_i}(w)=\emptyset$$

Which is impossible by Fact 5.3.14.

■

5.3.17. Corollary. *If the announcement of intention-rationality stabilizes at k for a given pointed model $\mathbb{M},w$, then $\mathbb{M}^k_{|\bigvee_{j\in I}IR_j},w\models\bigwedge_i \mathsf{i}_i\top$.*

With this in hand, we get the intended results almost automatically.

5.3.18. Fact. [Successfulness of $[\bigvee_{i\in I}IR_i!]$-stabilization] At any k where $[\bigvee_{i\in I}IR_i!]$ stabilizes, $\mathbb{M}^k_{|\bigvee_{i\in I}IR_i},w\models\bigvee_{i\in I}IR_i$.

Proof. Assume not, then $\mathbb{M}^k_{|\bigvee_{i\in I}IR_i},w\models\neg\bigvee_{i\in I}IR_i$. But then $w\not\in\mathbb{M}^{k+1}_{|\bigvee_{i\in I}IR_i}$, against the assumption that the announcement of $\bigvee_{i\in I}IR_i$ stabilizes at k. ■

This means that, even though non-epistemic cleaning is not necessarily safe after one announcement, it is in the long run. But the route to a stable contracted epistemic game frame is much quicker with epistemic cleaning[25].

5.3.19. Fact. For any pointed model $\mathbb{M},w$, the announcement of $\bigvee_{i\in I}IR^*_i$ stabilizes after one announcement.

Proof. By definition, $W_{|\bigvee_{i\in I}IR^*_i}=||\bigvee_{i\in I}IR^*_i||$. But we also know from Fact 5.3.12 that for all w' in $W_{|\bigvee_{i\in I}IR^*_i}$, $\mathbb{M}_{|\bigvee_{i\in I}IR^*_i},w'\models\bigvee_{i\in I}IR^*_i$. This means that $\mathbb{M}^2_{|\bigvee_{i\in I}IR^*_i}=\mathbb{M}_{|\bigvee_{i\in I}IR^*_i}$. ■

[25] The situation here is similar to what happens for announcements of weak and strong rationality in [van Benthem, 2003]. The comparison of these various transformations would certainly be illuminating. I look a briefly at their interaction in Section 5.3.3.

Moreover, as the example in the proof of Fact 5.3.13 suggests, these stabilization points can be slightly different.

5.3.20. Fact. [Fixed points divergence] There exist models $\mathbb{M}$ where the announcement of intention-rationality stabilizes at k such that :

$$\mathbb{M}_{|\bigvee_{i\in I} IR_i^*} \not\subseteq \mathbb{M}^k_{|\bigvee_{i\in I} IR_i}$$

Proof. Take a model $\mathbb{M}$ with two agents and four states, w_1 to w_4, where $[w]_i = \{w\}$ for all states. Fix the intentions as in Table 5.5. It should be clear that in all states, $\mathbb{M}, w \models \bigvee_{i\in I} IR_i$. This means that for all states, $\mathbb{M}_{|\bigvee_{i\in I} IR_i}, w = \mathbb{M}, w$, i.e. this announcement does not remove any states, and so that $\mathbb{M}$ is its own stabilization point. But observe, on the other hand, that at $\mathbb{M}, w_2 \not\models \bigvee_{i\in I} IR_i^*$. But since $\downarrow\iota_1(w_1) = \{w_2\}$, we get $\downarrow\iota_{1,|\bigvee_{i\in I} IR_i^*}(w_1) = \{w_1, w_3, w_4\}$ after the announcement of knowledge-consistency at w_1. But then it is clear that $\downarrow\iota_{1,|\bigvee_{i\in I} IR_i^*}(w_1) \not\subseteq \downarrow\iota_{1,|\bigvee_{i\in I} IR_i}(w_1)$, and since in this case the announcement of $\bigvee_{i\in I} IR_i$ "stabilizes" at $k = 0$, we get that $\mathbb{M}_{|\bigvee_{i\in I} IR_i^*} \not\subseteq \mathbb{M}^k_{\bigvee_{i\in I} IR_i}$ ■

w	$\downarrow\iota_1(w)$	$\downarrow\iota_2(w)$
w_1	w_2	w_1
w_2	w_1	w_4
w_3	w_3	w_3
w_4	w_4	w_4

Table 5.5: The intentions of the agents in counterexample for Fact 5.3.20.

This last result is essentially a consequence of the intention-revision policy. It preserves consistency of intentions, but it sometimes forces agents in epistemic game frames to adjust their intentions in the face of non-epistemic cleaning in a way that would not have been necessary for epistemic cleaning.

This difference between the fixed points of epistemic and non-epistemic cleaning is, however, the only one that can occur. In particular, knowledge-consistency is robust to any number of altruistic cleanings. This is, once again, something the agents can deduce in strategic games with intentions. If they are knowledge-consistent they can conclude that they remain so after the announcement of altruistic cleaning.

5.3.21. Fact. For all pointed models $\mathbb{M}, w$, if $\mathbb{M}, w \models \bigvee_{i\in I} IR_i^*$ then $\mathbb{M}, w \models [\bigvee_{i\in I} IR_i!] \bigvee_{i\in I} IR_i^*$.

Proof. Assume that $\mathbb{M}, w \models \bigvee_{i\in I} IR_i^*$, i.e. that there is an i and a $w' \sim_i w$ such that $w' \in \downarrow\iota_i(w)$. Because IR_i^* is introspective, this means that $\mathbb{M}, w' \models \bigvee_{i\in I} IR_i^*$. But then $\mathbb{M}, w' \models \bigvee_{i\in I} IR_i$, which means that both w' and w are in $W_{|\bigvee_{i\in I} IR_i}$, and also that $w' \in \downarrow\iota_{i|\bigvee_{i\in I} IR_i}(w)$. But then $\mathbb{M}, w \models [\bigvee_{i\in I} IR_i!] \bigvee_{i\in I} IR_i^*$. ■

5.3.22. Corollary. *Suppose that for a pointed model $\mathbb{M}, w$ the announcement of intention-rationality stabilizes at k, then*

$$\Big\|\bigvee_{i\in I} IR_i^*\Big\|_{|\bigvee_{i\in I} IR_i^*} \subseteq \Big\|\bigvee_{i\in I} IR_i\Big\|^k_{|\bigvee_{i\in I} IR_i}$$

Using $D\mathcal{L}_{\mathcal{GF}}$, we thus get a very general picture of cleaning-like operations and of their associated reasoning in epistemic game frames. Not only have I been able to recapture altruistic cleaning, but we have seen that there is in fact a whole family epistemic variants of this operation, which correspond to new epistemic criteria for admissibility. Much more could be said along these lines, of course. In particular, it would be interesting to have a systematic classification of the possible types of cleaning, according to their relative strengths or their long run behaviour. I do not, however, pursue these matters here. I rather look at two other aspects of model transformation involving intentions, namely the cases of intention overlap and of enabling announcements.

Intention overlap

We saw in Chapter 3 (Section 3.8) that overlap of intentions, or what I called intention agreement, is crucial to ensure coordination in the general case. In this section I look at whether intention overlap is something that could be forced, so to speak, by an arbitrary announcement in epistemic game frames. As we shall see, this is unfortunately not the case. The intentions of agents overlap after an arbitrary announcement only if they already overlapped before the announcement, at least for the agents whose intentions were compatible with the announcement. To see this, let me first fix the formal definition of intention overlap in epistemic game frames.

5.3.23. Definition. [Intention Overlap] At a pointed epistemic game frame $\mathbb{G}, w$, the intentions of the agents *overlap* whenever there is a $w' \in W$ such that $w' \in \downarrow\iota_i(w)$ for all $i \in I$.

Obviously, if the intentions of the agents overlap at a state w in a given epistemic game frame, then for any model based that frame we have: $\mathbb{M}, w \models \bigvee_{\sigma\in S}\bigwedge_{i\in I} \mathsf{i}_i\sigma$. With this in hand, one can directly show general conditions for overlapping.

5.3.24. Fact. [Intention overlap] For any pointed model $\mathbb{M}, w$, the following are equivalent:

(i) $\mathbb{M}, w \models [\phi!]\bigvee_{\sigma\in||\phi||}\bigwedge_{i\in I} \mathsf{i}_i\sigma$

(ii) There is a $w' \in ||\phi||$ such that for all $i \in I$, if $\downarrow\iota_i(w) \cap ||\phi|| \neq \emptyset$ then $w' \in\downarrow\iota_i(w)$.

Proof. The first part of the proof is be syntactical. I again use axioms from page 107 to "deconstruct" the post-announcement conditions of (i) into pre-announcement conditions. Then I show, via a correspondence argument, that these conditions are indeed those expressed by (ii). So let us start with:

$$[\phi!] \bigvee_{\sigma \in ||\phi||} \bigwedge_{i \in I} \mathsf{i}_i \sigma$$

This formula is propositionally equivalent to :

$$[\phi!] \neg \bigwedge_{\sigma \in ||\phi||} \neg \bigwedge_{i \in I} \mathsf{i}_i \sigma$$

Then, by (2) and (3) , we get:

$$\phi \to \neg \bigwedge_{\sigma \in ||\phi||} [\phi!] \neg \bigwedge_{i \in I} \mathsf{i}_i \sigma$$

By (2) again, we obtain:

$$\phi \to \neg \bigwedge_{\sigma \in ||\phi||} (\phi \to \neg \bigwedge_{i \in I} [\phi!] \mathsf{i}_i \sigma)$$

Now we can replace i_i by its dual:

$$\phi \to \neg \bigwedge_{\sigma \in ||\phi||} (\phi \to \neg \bigwedge_{i \in I} [\phi!] \neg I_i \neg \sigma)$$

We then reapply (2):

$$\phi \to \neg \bigwedge_{\sigma \in ||\phi||} (\phi \to \neg \bigwedge_{i \in I} (\phi \to [\phi] I_i \neg \sigma))$$

This, after some transformation from propositional logic, reduces to:

$$\phi \to \bigvee_{\sigma \in ||\phi||} \bigwedge_{i \in I} (\phi \to \neg [\phi!] I_i \neg \sigma)$$

Now I will look at $[\phi!] I_i \neg \sigma$ separately. By (5) it reduces to:

$$\phi \to ((\mathsf{i}_i \phi \wedge I_i(\phi \to [\phi!] \neg \sigma)) \vee (I_i \neg \phi \wedge [\phi!] A \neg \sigma))$$

By (2) and (4), this is the same as:

$$\phi \to ((\mathsf{i}_i \phi \wedge I_i(\phi \to \neg \sigma)) \vee (I_i \neg \phi \wedge A(\phi \to \neg \sigma))$$

Now, reinserting this formula in the main one and pushing the negation inside, we get:

$$\phi \to \bigvee_{\sigma \in ||\phi||} \bigwedge_{i \in I} (\phi \to (\mathsf{i}_i \phi \to \mathsf{i}_i(\phi \wedge \sigma)) \wedge (I_i \neg\phi \to E(\phi \wedge \sigma)))) \tag{Post}$$

I am now ready for the correspondence argument, which boils down to showing that (Post) is true at a pointed model $\mathbb{M}, w$ iff (ii) holds. We have that $\mathbb{M}, w, \models$ Post iff there is a σ and a $w' \in ||\phi||$ with $V(\sigma) = \{w'\}$ such that for all $i \in I$ the two conjuncts hold, provided that $\mathbb{M}, w \models \phi$. Observe first that the two conjuncts divide the agents into two groups. On the one hand are those such that $\downarrow\iota_i(w) \cap ||\phi|| \neq \emptyset$, i.e. those whose intentions are compatible with the announcement of ϕ. On the other hand are those whose intentions are not compatible with this announcement. Let us look at this case first, which is taken care of by the second conjunct $I_i \neg\phi \to E(\phi \wedge \sigma)$. This simply restates what we already knew, namely that ϕ holds at w', which means that the truth of this formula in fact only depends on the first conjunct, $\mathsf{i}_i \phi \to \mathsf{i}_i(\phi \wedge \sigma)$. This bluntly says that w' was already compatible with the intentions of all the agents in the first group before the announcement of ϕ, which is just what (ii) enforces. ■

This result tells us that intention overlap occurs after an arbitrary truthful announcement only in cases where it already occurred before, at least for the agents that had intentions compatible with the announcement. In other words, announcing arbitrary truths is not sufficient to force intention overlap. This is also something which features in the agents' reasoning about epistemic game frames, and the proof of this in fact explicitly shows part of this reasoning.

To ensure intention-overlap, and with it coordination in the general case, one has to look for more specific, i.e. stronger, forms of announcement. For example, the blunt announcement of the current strategy profile would surely do it. But this is not a very interesting case, for agents typically do not know what the actual profile is in epistemic game states. It would be more interesting to find introspective formulas of $\mathcal{L}_{\mathcal{GF}}$ whose announcement would ensure intention overlap.

Fact 5.3.24 states that overlap occurs after an announcement whenever it occurred before, *for the agents whose intentions were compatible with the announcement*. This means that, for the others, arbitrary announcements do force overlap. In fact, they do so in a very blunt way. All agents which have intentions inconsistent with a given announcement end up with the same intention set in the contracted epistemic game frame. So far, I have neglected this important aspect of the intention revision policy. It revises the intentions of agents in a uniform manner, brushing aside all differences in the pre-announcement intentions of such agents. From that point of view, the revision policy that I embodied here does indeed appear coarse-grained. This can be taken as a good reason to seek a more subtle policy of intention revision. But this can also be viewed from a more positive point of view. Precisely because it is so "blunt", the intention revision policy I use in this chapter also ensures intention overlap in case of revision. In other words, together with the coordination results in Chapter 3, it brings together the reasoning-centered and the volitive commitment of intentions.

Conditions for the enablement of cleaning

So far my attention has mainly focused on cases where cleaning does remove states from the epistemic game frame. But it can well be that this operation leaves the structure unchanged. In such cases, cleaning might benefit from other announcements to get started, so to speak. Here I provide one case study for such "enabling announcements" [van Benthem, 2003]. I look at the conditions under which announcing weak rationality (Section 5.2.2) enables epistemic cleaning. Let me first explain formally what enabling announcements are.

5.3.25. Definition. [Enabling announcements] The announcement of ϕ *enables* the announcement of ψ for a given model $\mathbb{M}$ whenever the following holds:

- $\mathbb{M}_{|\psi} = \mathbb{M}$
- $\mathbb{M}_{|\phi_{|\psi}} \subset \mathbb{M}_{|\phi} \subset \mathbb{M}$

With non-empty W for all these models.

In words, an announcement is enabling of another whenever, on the one hand, the second would by itself leave the original model unchanged but, on the other hand, it does change the model after announcement of the first. In other words, an announcement of ϕ enables the announcement of ψ when information in an epistemic game frame is not affected by the announcement of ψ alone, but it is affected by this announcement after the announcement of ϕ has taken place. As one can expect, announcing weak rationality enables cleaning under specific conditions at the interplay between what agents intend and what they prefer.

5.3.26. Fact. [Enabling IR_i^* with WR_j] For any model $\mathbb{M}$, the announcement of WR_i enables the announcement of IR_j^* iff for all w, $\mathbb{M}, w \models IR_j$ but there are some w such that

- $\mathbb{M}, w \models WR_i$.
- $\downarrow\iota_j(w) \cap ||WR_i|| \neq \emptyset$.
- for all $w' \in \downarrow\iota_j(w) \cap [w]_j$, $\mathbb{M}, w' \not\models WR_i$.

Proof. First the left-to-right direction. Assume that in $\mathbb{M}$ the announcement of WR_i enables the announcement of IR_j. That means first that $\mathbb{M}_{|IR_j} = \mathbb{M}$, which means that for all $w \in W$, $\mathbb{M}, w \models IR_j$, i.e. $\downarrow\iota_j(w) \cap [w]_j \neq \emptyset$. On the other hand, that $\mathbb{M}_{|WR_{i_{|IR^*_j}}} \subset \mathbb{M}_{|WR_i}$ means that for some of these w, we have $\mathbb{M}, w \models WR_i$ but $\downarrow \iota_{j,|WR_i}(w) \cap [w]_{j|WR_i} = \emptyset$. Now $\iota_{j,|WR_i}(w)$ can be obtained in two ways, depending on whether $\downarrow\iota_j(w) \cap ||WR_i|| = \emptyset$ or not. Consider the first case. We then would have $\iota_{j,|WR_i}(w) = \{||WR_i||\}$. But given that $||WR_i||$ is not empty for any game model, see [van Benthem, 2003, p.17] and that $[w]_{j|WR_i} \subseteq ||WR_i||$, we would conclude against our assumption that $\downarrow\iota_{j,|WR_i}(w) \cap [w]_{j|WR_i} \neq \emptyset$. So it has to be that $\downarrow\iota_j(w) \cap ||WR_i|| \neq \emptyset$. In this case $\iota_{j,|WR_i}(w) = \uparrow(\downarrow\iota_j(w) \cap ||WR_i||)$. This means that there are some $w' \in \downarrow\iota_j(w) \cap [w]_i$ while $w' \notin \downarrow\iota_{j|WR_i}(w) \cap [w]_{j|WR_j}$. Now, unpacking the definitions of the update rule for the intention set and the epistemic relation, we get, for all w:

$$\text{if } w \in (\downarrow\iota_i(w) \cap [w]_i \cap ||WR_j||) \text{ then } w \in \downarrow\iota_{i|WR_j}(w) \cap [w]_{i|WR_j}$$

Putting this in contrapositive, we get that for all $w' \notin \downarrow\iota_{j|WR_i}(w) \cap [w]_{j|WR_i}$ while being in $\downarrow\iota_j(w) \cap [w]_j$, $w' \notin ||WR_i||$

For the right-to-left direction, observe first that for all w, $\mathbb{M}, w \models IR_j$ is the same as to say that $\mathbb{M}_{|IR_j} = \mathbb{M}$. Now, take one w as specified. Since $\mathbb{M}, w \models WR_i$ and $\mathbb{M}, w' \not\models WR_i$ for all $w' \in \downarrow\iota_j(w) \cap [w]_j$, we know that $\mathbb{M}_{|WR_i} \subset \mathbb{M}$ and that the former is not empty. Now, because $\downarrow\iota_j(w) \cap ||WR_i|| \neq \emptyset$ we also know that $\iota_{j,|WR_i}(w) = \uparrow(\downarrow\iota_j(w) \cap ||WR_i||)$, which means that $\downarrow\iota_{j,|WR_i}(w) \subseteq \downarrow\iota_j(w)$. Moreover $[w]_{j|WR_i} \subseteq [w]_j$ by definition. This means that $([w]_{j|WR_i} \cap \downarrow\iota_{j,|WR_i}(w)) \subseteq ([w]_j \cap \downarrow \iota_j(w))$. What is more, by assumption, $(\downarrow\iota_j(w) \cap [w]_j) \subseteq ||\neg WR_i||$, which in the present context can only result in $\downarrow \iota_{j|WR_i}(w) \cap [w]_{j|WR_i} = \emptyset$. This means that $w \notin W_{|WR_{i_{|IR^*_j}}}$, and so that $\mathbb{M}_{|WR_{i_{|IR^*_j}}} \subset \mathbb{M}_{|WR_i}$. ■

This fact shows that cleaning is enabled by announcement of weak rationality just in case some agents have formed intentions without taking the rationality of others into account. They intend to achieve profiles that their opponent would never rationally choose. This is of course reminiscent of the interplay we saw in Chapter 4 between pruning and preferences in cases of sub-optimal picking functions (Section 4.2). What is going on here is indeed that weak-rationality enables cleaning only when the agents did not take into account that they are interacting with other rational agents when forming their intentions. They count on the others, as it were, to play irrational strategies in order to achieve their intentions.

Here the interplay between agents is in fact crucial. Because weak rationality is also introspective[26], agents cannot "enable themselves" by first announcing their own rationality and then their own knowledge consistency.

[26] See again van Benthem [2003] for a proof of this fact.

5.3.27. Fact. [No self-enabling] If $\mathbb{M}, w \models WR_j$ and $\mathbb{M}, w' \models IR_i^*$ for all $w' \in W$, and WR_j enables IR_i^*, then $i \neq j$.

Proof. We would be done if we can show that $\mathbb{M}_{|WR_{i_{|IR_i^*}}} = \mathbb{M}_{|WR_i}$ provided that $\mathbb{M}, w \models WR_i$ and $\mathbb{M}, w' \models IR_i^*$ for all $w' \in W$. This, in turn, follows from the Lemma 5.3.28 (below). ■

5.3.28. Lemma. *For any pointed model* $\mathbb{M}, w$, *if* $\mathbb{M}, w \models IR_i^*$ *and* $\mathbb{M}, w \models WR_i$ *then* $\mathbb{M}, w \models [WR_i!]IR_i^*$.

Proof. I will show the contrapositive. The proof rests crucially on the fact that WR_i is introspective. That is, for any $\mathbb{M}, w \models WR_i$ we also have $\mathbb{M}, w' \models WR_i$ for all $w' \sim_i w$.

Assume that $\mathbb{M}, w \not\models [WR_i!]IR_i^*$ while $\mathbb{M}, w \models WR_i$. That means that $\mathbb{M}_{|WR_i}, w \not\models IR_i^*$, i.e. that $\downarrow\iota_{i|WR_i}(w) \cap [w]_{i|WR_i} = \emptyset$. Now, as usual, $\iota_{i|WR_i}(w)$ can come from two sources.

1. It can be that $\iota_{i|WR_i}(w) = \{W_{|WR_i}\}$ because $\downarrow\iota_i(w) \cap ||WR_i|| = \emptyset$. But because $\mathbb{M}, w \models WR_i$ and WR_i is introspective, we know that $[w]_i \subseteq ||WR_i||$, which means that $\downarrow\iota_i(w) \cap [w]_i = \emptyset$, i.e. $\mathbb{M}, w, \not\models IR_i^*$.

2. It thus remains to check what happens when $\downarrow\iota_i(w) \cap ||WR_i|| \neq \emptyset$. I will show that in that case $\downarrow\iota_i(w) \cap [w]_i$ and $\downarrow\iota_{i|WR_i}(w) \cap [w]_{i|WR_i}$ are just the same set.

   The right-to-left inclusion follows directly from the definitions of the restricted relations and intention set. Now take a $w' \in\downarrow \iota_i(w) \ \cap \ [w]_i$. Since $w' \sim_i w$ and WR_i is introspective we know that $w' \in W_{|WR_i}$ and $w' \sim_{i|WR_i} w$. But since $\downarrow \iota_i(w) \cap ||WR_i|| \neq \emptyset$, we also know that $\downarrow \iota_{i|WR_i}(w) =\downarrow\iota_i(w) \cap ||WR_i||$, which means that w' is in $\downarrow\iota_{i,|WR_i}(w)$ as well.

   So $\downarrow\iota_i(w) \cap [w]_i =\downarrow\iota_{i|WR_i}(w) \cap [w]_{i|WR_i}$. The required conclusion is then a direct consequence of our assumption that $\downarrow\iota_{i|WR_i}(w) \cap [w]_{i|WR_i} = \emptyset$.

■

This proof once again displays a valid principle of the logic of strategic games with intentions: knowledge-consistency is robust to the announcement of weak rationality. This means that this is also something that planning agents can deduce in game situations. This exemplifies very well the kind of intention-based practical reasoning that the present logic can provide: a reasoning precisely at the intersection of instrumental rationality and planning agency.

Looking more systematically at enabling announcements can surely contribute to our general understanding of intention-based transformations of decision problem. As van Benthem [2003] suggests, one might also find interesting cases where intention-based announcements enable weak rationality. Given that the latter correspond to the well-known game theoretical process of elimination of strongly dominated strategies, this would open the door to a nice interplay between intention-based and classical game-theoretical reasonings. I shall not pursue that here, however. In Appendix 5.5.6 I show, much in the spirit of van Benthem [2003], that Nash equilibria can be given a dynamic characterization in $D\mathcal{L}_{\mathcal{GF}}$. This already indicates that this language, and with it the whole "DEL methodology", is also quite suited to capturing game-theoretical notions.

These consideration close the section on the dynamics of epistemic game frames with intention. Before looking back at the whole chapter, let me briefly review what we saw in this section.

By taking a logical stance, I have connected cleaning to the notion of intention-rationality and I have situated it within a bigger family of option-excluding operations. In particular, I studied the connection between two forms of cleaning, the altruistic version and its "epistemic" variant. I showed that these operations behave quite similarly in epistemic game frames with state-dependent intentions, but that they might diverge in the long run once this assumption is lifted. As I noted, the coarse-grained intention revision policy that I introduced is mostly responsible for this divergence. I gained a better assessment of the pros and cons of this policy by taking a look at condition for intention overlap. I also investigated the interplay between cleaning and announcement of weak rationality, and provided conditions under which the second enables the first. All through this section, I also pointed to many instances of valid principles in epistemic games frames which, by the completeness result of Section 5.3.2, correspond directly to reasoning of planning agents in such interactive situations.

5.4 Conclusion

In this chapter, I have proposed a unified theory of practical reasoning in interactive situations with intentions. We have seen that some aspects of the volitive commitment of intentions echo their reasoning-centered commitments, e.g. intention-rationality and exclusion of inadmissible options. I have also been able to match conditions on what the agents know and intend with epistemic game frame transformations, e.g. knowledge consistency and "epistemic" cleaning. Taking the logical point of view also allowed me to venture into new territories, namely policies of intention-revision, general conditions for overlap of intentions and enabling of model transformation, all provided with a concrete deductive counterpart. Even though there is still a lot to explore about these three topics, I hope to have open the way towards a fully developed theory of intention-based practical reasoning in games.

Throughout this chapter, and more generally in this thesis, I have made no attempt to connect with another important paradigm for intention-based practical reasoning, the so-called BDI (Belief-Desire-Intention) architectures for multi-agent systems[27]. Although very similar in method and aims, the BDI models have not been developed for direct application to strategic games, which makes it at least not trivial to see how they relate to the present framework. It would nevertheless be worthwhile investigating the connection. The work on intention revision of van der Hoek et al. [2007], which is strongly based on the BDI architectures, can definitely enrich what I have proposed here, and the explicit focus on games could arguably profit BDI reasoning, too. I shall not, however, go in that direction in the next chapter. I rather take a step back and ask why, to start with, intentions play such a role for planning agents. This will allow me to clarify some philosophical concepts that I used since the beginning of this thesis, while at the same time opening the door to unexpected avenues for practical reasoning with intentions in strategic interaction.

5.5 Appendix

5.5.1 Bisimulation and modal equivalence for $\mathcal{L}_\mathcal{P}$

5.5.1. Definition. [Modal equivalence] Two pointed preference models $\mathbb{M}, w$ and $\mathbb{M}', v$ are *modally equivalent*, noted $\mathbb{M}, w \leftrightsquigarrow \mathbb{M}', v$, iff for all formula ϕ of $\mathcal{L}_\mathcal{P}$, $\mathbb{M}, w \models \phi$ iff $\mathbb{M}', v \models \phi$.

5.5.2. Definition. [Bisimulation] Two pointed preference models $\mathbb{M}, w$ and $\mathbb{M}', v$ are *bisimilar*, noted $\mathbb{M}, w \underline{\leftrightarrow} \mathbb{M}', v$, whenever there is a relation $E \subseteq \mathbb{M} \times \mathbb{M}'$ such that:

1. For all $p \in \text{PROP}, w \in V(p)$ iff $v \in V(p)$,
2. (Forth) if $w' \succeq w$ ($w' \succ w$) then there is a $v' \in W'$ such that $v' \succeq' v$ ($v' \succ' v$ respectively) and $w'Ev'$,
3. (Back) if $v' \succeq' v$ ($v' \succ' v$) then there is a $w' \in W$ such that $w' \succeq w$ ($w' \succ w$ respectively) and $v'Ew'$,
4. For all $w' \in W$, there is a $v' \in W'$ such that $w'Ev'$, and
5. For all $v' \in W'$, there is a $w' \in W$ such that $v'Ew'$.

The relation E is called a *total bisimulation* between $\mathbb{M}, w$ and $\mathbb{M}', v$. If E is a bisimulation and $\mathbb{M}, w \underline{\leftrightarrow} \mathbb{M}', v$, then we say that w and v are bisimilar, which is noted $w \underline{\leftrightarrow} v$.

[27] Key references here are Cohen and Levesque [1990], Georgeff et al. [1998] and Wooldridge [2000].

As usual in modal logic, one can show that any two bisimilar pointed preference models are modally equivalent. In other words, truth in $\mathcal{L}_{\mathcal{P}}$ is invariant under bisimulation.

5.5.2 More on lifted relations in $\mathcal{L}_{\mathcal{P}}$

The following lifted preference relations can be defined in $\mathcal{L}_{\mathcal{P}}$:

5.5.3. Definition. [Binary preference statements]

$$\begin{array}{lrcl} 1. & \psi \geq_{\exists\exists} \phi & \Leftrightarrow & E(\phi \wedge \Diamond^{\leq}\psi) \\ 2. & \phi \leq_{\forall\exists} \psi & \Leftrightarrow & A(\phi \rightarrow \Diamond^{\leq}\psi) \\ 3. & \psi >_{\exists\exists} \phi & \Leftrightarrow & E(\phi \wedge \Diamond^{<}\psi) \\ 4. & \phi <_{\forall\exists} \psi & \Leftrightarrow & A(\phi \rightarrow \Diamond^{<}\psi) \end{array}$$

The formulas $\psi \geq_{\exists\exists} \phi$ and $\psi >_{\exists\exists} \phi$ may be read as "there is a ψ-state that is at least as good as a ϕ-state" and "there is a ψ-state that is strictly better than a ϕ-state", respectively. The other comparative statements, $\phi \leq_{\forall\exists} \psi$ and $\phi <_{\forall\exists} \psi$, can be read as "for all ϕ-states there is an at least as good ψ-state" and as "for all ϕ-state there is a strictly preferred ψ-state", respectively.

We can define further binary preference statements as duals of the above modalities.

5.5.4. Definition. [Duals]

$$\begin{array}{lrcccl} 5. & \phi >_{\forall\forall} \psi & \Leftrightarrow & \neg(\psi \geq_{\exists\exists} \phi) & \Leftrightarrow & A(\phi \rightarrow \Box^{\leq}\neg\psi) \\ 6. & \phi >_{\exists\forall} \psi & \Leftrightarrow & \neg(\phi \leq_{\forall\exists} \psi) & \Leftrightarrow & E(\phi \wedge \Box^{\leq}\neg\psi) \\ 7. & \phi \geq_{\forall\forall} \psi & \Leftrightarrow & \neg(\psi >_{\exists\exists} \phi) & \Leftrightarrow & A(\phi \rightarrow \Box^{<}\neg\psi) \\ 8. & \phi \geq_{\exists\forall} \psi & \Leftrightarrow & \neg(\phi <_{\forall\exists} \psi) & \Leftrightarrow & E(\phi \wedge \Box^{<}\neg\psi) \end{array}$$

The first formula tells us that "everywhere in the model, if ϕ is true at a state, then there is no ψ-state at least as good as it". Observe that if the underlying preference relation is total, this boils down to saying that all ϕ-states, if any, are strictly preferred to all ψ-states, also if any. This is indeed the intended meaning of the notation $\phi >_{\forall\forall} \psi$. Similarly, the second dual says, under assumption of totality, that there is a ϕ-state strictly preferred to all the ψ-states, if any. These intended meaning are, however, not definable in $\mathcal{L}_{\mathcal{P}}$ without assuming totality.

5.5.5. Fact. The connectives $\phi >_{\forall\forall} \psi$, $\phi >_{\exists\forall} \psi$, $\phi \geq_{\forall\forall} \psi$ and $\phi \geq_{\exists\forall} \psi$, in their intended meaning, are not definable in $\mathcal{L}_{\mathcal{P}}$ on non-totally ordered preference frames.

Proof. See van Benthem et al. [Forthcoming]. ■

5.5.6. Definition. [Relation Lifting] A property *lifts* with a relation $\leq$ in the class of models M if whenever $\succeq$ as this property the lifted relation $\leq$ also has it.

5.5.7. Fact. With respect to the class of preference models, reflexivity and totality lift with $\geq_{\exists\exists}$ for satisfied formulas and with $\leq_{\forall\exists}$ for any formulas. Transitivity does lift with $\leq_{\forall\exists}$ for satisfied formulas, but not with $\geq_{\exists\exists}$.

Proof. The proof is trivial for reflexivity, in both cases, for totality with $\geq_{\exists\exists}$ and for transitivity with $\leq_{\forall\exists}$, both for satisfied formulas. Totality for $\leq_{\forall\exists}$ has been proved in Section 5.1.2.

For failure of transitivity lifting with $\geq_{\exists\exists}$, take a preference model with four states where $w_1 \succeq w_2 \succeq w_3 \succeq w_2$. Make ϕ only true at w_1 and w_4, ψ only at w_3 and ξ only at w_3. We clearly get $\psi \geq_{\exists\exists} \phi$, $\phi \geq_{\exists\exists} \xi$ but not $\psi \geq_{\exists\exists} \xi$. ■

Halpern [1997] and Liu [2008] have other results of this kind, with different binary relations among formulas.

5.5.3 More on the expressive power of $\mathcal{L}_{\mathcal{GF}}$

In the following definition, the clauses 2 and 3 are intended to apply to both $\succeq_i$ and $\sim_i$. For that reason I simply write R.

5.5.8. Definition. [Bisimulation] Two pointed game pointed models $\mathbb{M}, w$ and $\mathbb{M}', v$ are *bisimilar*, noted $\mathbb{M}, w \leftrightarroweq \mathbb{M}', v$, whenever there is a relation $E \subseteq \mathbb{M} \times \mathbb{M}'$ such that:

1. For all $x \in \textsc{prop} \cup S, w \in V(x)$ iff $v \in V(x)$,
2. (Forth -R) if $w' \, R \, w$ then there is a $v' \in W'$ such that $v' \, R \, v$ and $w'Ev'$,
3. (Back -R) if $v' \, R \, v$ then there is a $w' \in W$ such that $w' \, R \, w$ and $v'Ew'$,
4. [ten Cate, 2005, p.47] For all $\sigma \in S$, if $V(\sigma) = \{w\}$ and $V'(\sigma) = \{v\}$ then wEv.
5. [Hansen, 2003, p.18] (Forth -ι) if $X \in \iota_i(w)$ then there is a $X' \subseteq W'$ such that $X' \in \iota(v)$ and for all $v' \in X'$ there is a $w' \in X$ such that $w'Ev'$.
6. [Hansen, 2003, p.18] (Back -ι) if $X' \in \iota_i(v)$ then there is a $X \subseteq W$ such that $X \in \iota(w)$ and for all $w' \in X'$ there is a $v' \in X$ such that $v'Ew'$.

Truth in $\mathcal{L}_{\mathcal{GF}}$ is indeed invariant under this notion of bisimulation.

5.5.9. Fact. [Nash Equilibrium without nominals] Let $\mathcal{L}_{\mathcal{GF}}{}^-$ be $\mathcal{L}_{\mathcal{GF}}$ minus the nominals. Nash equilibrium is not definable in $\mathcal{L}_{\mathcal{GF}}{}^-$.

G_1	t_1	t_2
s_1	1	0
s_2	1	0

G_2	t_0	t_1	t_2
s_1	3	1	0
s_2	2	1	0

Table 5.6: Two games with bisimilar models but different Nash equilibria. The payoffs are identical for both players.

Proof. Look at the pair of games in Table 5.6, where the payoffs are the same for both agents. Take the models for these games depicted in Figure 5.10. Assume that for all $w \in W$ and $i \in \{1, 2\}$, $\iota_i(w) = \{W\}$, $\{w' : w \sim_i w'\} = \{w\}$ and for all $p \in prop$, $V(p) = \emptyset$, and similarly in $\mathbb{M}'$.

It is easy to check that $\mathcal{L}_{\mathcal{GF}}{}^{-}$ is invariant under bisimulation as just defined, without clauses related to nominals. $\mathbb{M}$ and $\mathbb{M}'$ are bisimilar in that sense, as reveals a rather tedious check from the information in Table 5.7. Now observe that v_3 is a Nash equilibrium, while one of its bisimilar counterpart, w_3, is not.

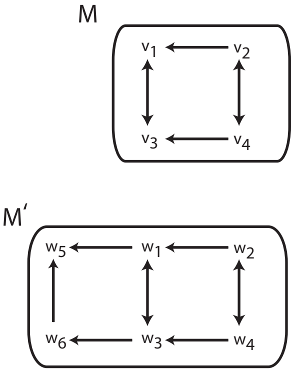

Figure 5.10: The models for the games in Table 5.6. Only the preference relations are represented.

■

5.5.4 Proof of Theorem 5.2.13

The proof is essentially a collage of known techniques for the various fragments of $\mathcal{L}_{\mathcal{GF}}$. Before going into detail, let me give a brief survey of the main steps.

The first part amounts to ensuring that we can build a *named* model for any consistent set of formulas in $\mathcal{L}_{\mathcal{GF}}$. A named model is a model where all states are indeed “named” by at least one nominal in the language. Once this is secured, we can really profit from the expressive power provided by nominals. In such models all properties definable by a *pure formula*, i.e. a formula with only nominals as atoms, are cannonical (see ten Cate [2005, p.69]). During the construction of

Profile in G_1	State in M	bisimilar to in M'
(s_1, t_1)	v_1	w_1, w_2, w_3, w_5, w_6
(s_1, t_2)	v_2	w_2
(s_2, t_1)	v_3	w_1, w_3, w_5
(s_2, t_2)	v_4	w_4

Profile in G_2	State in M'
(s_1, t_1)	w_1
(s_1, t_2)	w_2
(s_2, t_1)	w_3
(s_2, t_2)	w_4
(s_1, t_0)	w_5
(s_2, t_0)	w_6

Table 5.7: The bisimulation for the models in Figure 5.10.

the named model we also make sure that it contains enough states to prove an existence lemma for E, which is a little trickier than usual in the presence of nominals. This boils down to showing that it is *pasted*, a property that is defined below. All this is routine for hybrid logic completeness. Most definitions and lemmas come from Blackburn et al. [2001, p.434-445] and Gargov and Goranko [1993].

I then turn to the other fragments of $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$, by proving existence lemmas for K_i, $\Diamond^{<}$, $\Diamond^{\leq}$ and I_i. These are completely standard, just like the truth lemma that comes thereafter. In the only part of the proof that is specific to $\mathcal{L}_{\mathcal{GF}}$, I finally make sure that the model can be seen as an epistemic game model with intentions. As we shall see, this is a more or less direct consequence of known facts about neighbourhood semantics, see [Pacuit, 2007], together with the aforementioned canonicity of pure formulas and the various interaction axioms. From this we will have shown completeness with respect to the class of epistemic game models with intentions.

5.5.10. Definition. [Named and pasted MCS] Let Γ be a maximally consistent set (MCS) of $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$. We say that Γ is *named* by σ if $\sigma \in \Gamma$. If σ names some MSC(s) Γ we denote it (them) Γ_σ. Γ is *pasted* whenever $E(\sigma \wedge <> \phi) \in \Gamma$ implies that $E(\sigma \wedge <> \sigma') \wedge E(\sigma' \wedge \phi)$ is also in Γ.

5.5.11. Lemma (Extended Lindenbaum lemma). *[Blackburn et al., 2001, p.441] Let S' be a countable collection of nominals disjoint from S, and let $\mathcal{L}_{\mathcal{GF}}{}'$ be $\mathcal{L}_{\mathcal{GF}} \cup S'$. Then every $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$ consistent set of formulas can be extended to a named and pasted $\Lambda_{\mathcal{L}_{\mathcal{GF}}'}$-MCS.*

Proof.

Naming Enumerate S', and let σ be the first new nominal in that enumeration. For a given consistent set Γ^*, fix $\Gamma_\sigma = \Gamma \cup \{\sigma\}$. By (Name) Γ_σ is consistent.

Pasting Enumerate the formulas of $\mathcal{L}_{\mathcal{GF}}{}'$ and take $\Gamma_0 = \Gamma_\sigma$. Assume Γ_n is defined, and let ϕ_{n+1} be the $n^{th}+1$ formula in the enumeration. Define Γ_{n+1} as Γ_n if $\Gamma_n \cup \{\phi_{n+1}\}$ is inconsistent. Otherwise form Γ_{n+1} by adding ϕ_{n+1} to Γ_n if ϕ_{n+1} is not of the form $E(\sigma' \wedge \psi)$. If ϕ_{n+1} is of form $E(\sigma' \wedge \psi)$, then we paste with the first new nominal σ'' in the enumeration of S'. I.e. $\Gamma_{n+1} = \Gamma_n \cup \{\phi_{n+1}\} \cup \{E(\sigma' \wedge <> \sigma'') \wedge E(\sigma'' \wedge \phi)\}$. By (Paste), Γ_{n+1} is also consistent. Set, finally, $\Gamma = \bigcup_{n \leq \omega} \Gamma_n$. This is clearly a named and pasted MCS.

■

5.5.12. Definition. [Yielded MCS] The sets *yielded* by a $\Lambda_{\mathcal{L}_{\mathcal{GF}}{}'}$-MCS Γ are the sets Δ_σ such that $\Delta_\sigma = \{\phi : E(\sigma \wedge \phi) \in \Gamma\}$.

5.5.13. Lemma (Properties of yielded sets). *[Blackburn et al., 2001, p.439] Let Δ_σ and $\Delta_{\sigma'}$ be any yielded sets of a $\Lambda_{\mathcal{L}_{\mathcal{GF}}{}'}$-MCS Γ, for arbitrary nominals σ and σ' in $\mathcal{L}_{\mathcal{GF}}{}'$.*

1. *Both Δ_σ and $\Delta_{\sigma'}$ are named $\Lambda'_{\mathcal{L}_{\mathcal{GF}}}$-MCS.*

2. *If $\sigma' \in \Delta_\sigma$ then $\Delta_\sigma = \Delta_{\sigma'}$.*

3. *$E(\sigma \wedge \phi) \in \Delta_{\sigma'}$ iff $E(\sigma \wedge \phi) \in \Gamma$.*

4. *If σ'' names Γ then Γ is itself the yielded set $\Delta_{\sigma''}$.*

Proof.

1. By (Exists$_\sigma$), $E\sigma \in \Gamma$, and thus Δ_σ is named. Assume now it is not consistent. That means that there are $\xi_1 \wedge ... \wedge \xi_n$ such that one can prove $\neg(\xi_1 \wedge ... \wedge \xi_n)$ in $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$. But that means that $A\neg(\xi_1 \wedge ... \wedge \xi_n) \in \Gamma$, by (Nec). This, in turns, means that $\neg E(\xi_1 \wedge ... \wedge \xi_n) \in \Gamma$. But that can't be. Recall that $(\xi_1 \wedge ... \wedge \xi_n) \in \Delta_\Gamma$ iff $E(\sigma \wedge \xi_1 \wedge ... \wedge \xi_n)$ is also in Γ. But then by (K) for E, we get that $E(\xi_1 \wedge ... \wedge \xi_n) \in \Gamma$. For maximality, observe that a formula ϕ and its negation are not in Δ_σ iff neither $E(\sigma \wedge \phi)$ nor $E(\sigma \wedge \neg\phi)$ are in Γ. But because the latter is a MCS, that means that both $\neg E(\sigma \wedge \phi)$ and $\neg E(\sigma \wedge \neg\phi)$ are in Γ. The first formula implies $A(\sigma \rightarrow \neg\phi) \in \Gamma$, but then, given that $E\sigma \in \Gamma$, by a standard modal logic reasoning we get that $E(\sigma \wedge \neg\phi)$, contradicting the consistency of Γ.

2. Assume $\sigma' \in \Delta_\sigma$. That means that $E(\sigma \wedge \sigma') \in \Gamma$. By ($\text{Inc}_{E-\sigma}$) we get that both $A(\sigma \to \sigma')$ and $A(\sigma' \to \sigma)$ are in Γ, and so by K for E, we get $A(\sigma \leftrightarrow \sigma') \in \Gamma$. Assume now that $\phi \in \Delta_\sigma$. This means that $E(\sigma \wedge \phi) \in \Gamma$. But standard K reasoning we get that $E(\sigma' \wedge \phi) \in \Gamma$, which means that ϕ is also in $\Delta_{\sigma'}$. The argument is symmetric for $\phi \in \Delta_{\sigma'}$, and so $\Delta_\sigma = \Delta_{\sigma'}$.

3. I first show the left-to-right direction. Assume that $E(\sigma' \wedge \phi) \in \Delta_\sigma$. This means that $E(\sigma \wedge E(\sigma' \wedge \phi)) \in \Gamma$. But then this implies, by K for E that $EE(\sigma' \wedge \phi) \in \Gamma$, which in turns, because of (Trans) for E, implies $E(\sigma' \wedge \phi) \in \Gamma$. For the converse, assume that $E(\sigma' \wedge \phi) \in \Gamma$. By (Sym) for E, we get that $AE(\sigma' \wedge \phi) \in \Gamma$. But we also know by ($\text{Exists}_\sigma$) that $E\sigma \in \Gamma$, from which we get by standard K reasoning that $E(\sigma \wedge E(\sigma' \wedge \phi)) \in \Gamma$. This means that $E(\sigma' \wedge \phi) \in \Delta_\sigma$.

4. Assume that $\sigma \in \Gamma$. For the left to right, assume that $\phi \in \Gamma$. This means that $\sigma \wedge \phi \in \Gamma$, which implies by (Ref) that $E(\sigma \wedge \phi)$ and so that $\phi \in \Delta_\sigma$. Now assume that $\phi \in \Delta_\sigma$. This means that $E(\sigma \wedge \phi) \in \Gamma$, which in turn implies that $A(\sigma \to \phi)$ by ($\text{Inc}_{E-\sigma}$). But then by (Ref) again we get that $\sigma \to \phi \in \Gamma$, and ϕ itself because $\sigma \in \Gamma$.

■

This last lemma prepares the ground for the hybrid fragment for $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$. Now we need a few more background notions regarding the neighborhood fragment for this logic.

5.5.14. Definition. [Varieties of neighbourhood functions] [Pacuit, 2007, p.8-9] For any set W, we say that $f : W \to \mathcal{P}(\mathcal{P}(W))$ is:

- *closed under supersets* provided that for all w and each $X \in f(w)$, if $X \subseteq Y \subseteq W$ then $Y \in f(w)$.
- *closed under binary intersections* provided that for all w and each $X, Y \in f(w)$, $X \cap Y$ is also in $f(w)$.
- *a filter* if it is closed under supersets and under binary intersection.

5.5.15. Definition. [Neighbourhood tools] [Pacuit, 2007, p.8-9] Let W^Γ be the set of all named sets yielded by Γ.

- The *proof set* $|\phi|$ of a formula ϕ of $\mathcal{L}_{\mathcal{GF}}{}'$ is defined as $\{\Delta_\sigma \in W^\Gamma : \phi \in \Delta_\sigma\}$.
- A neighbourhood function ι is *canonical* for $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$ if for all ϕ, $|\phi| \in \iota(\Delta_\sigma)$ iff $I_i\phi \in \Delta_\sigma$.
- A neighbourhood function $\iota_{min} : W^\Gamma \to \mathcal{P}(\mathcal{P}(W))$ is *minimal* if $\iota_{min}(\Delta_\sigma) = \{|\phi| : \phi \in \Delta_\sigma\}$.

- The *supplementation* $\uparrow\iota_{min}$ of ι_{min} is the smallest function that contains $\iota_{min}(\Delta_\sigma)$ and that is closed under supersets.

5.5.16. Fact. [Properties of $\uparrow\iota_{min}$] [Pacuit, 2007] $\uparrow\iota_{min}$ is well-defined, canonical for $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$ and a filter.

5.5.17. Definition. [Epistemic model for completeness] Let Γ be any named and pasted $\Lambda_{\mathcal{L}_{\mathcal{GF}}'}$-MCS. The named game model $\mathbb{M}^\Gamma$ yielded by Γ is a tuple $\langle\ W^\Gamma, I, \sim_i^\Gamma, \succeq_i^\Gamma, \succ_i^\Gamma, \iota_i^\Gamma, V^\Gamma\ \rangle$ such that:

- W^Γ is the set of sets yielded by Γ.
- I, defined as $\{i :$ there is a $<>_i \phi$ in $\mathcal{L}_{\mathcal{GF}}\}$, is the set of agents.
- $\Delta_\sigma \sim_i^\Gamma \Delta_{\sigma'}$ iff for all $\phi \in \Delta_{\sigma'}$, $\Diamond_i\phi \in \Delta_\sigma$, and similarly for $\succeq_i^\Gamma$ and $\succ_i^\Gamma$.
- $\iota_i^\Gamma(\Delta_\sigma) = \uparrow\iota_{i,min}^\Gamma$.
- For all $x \in \text{PROP} \cup (S \cup S')$, $V^\Gamma(x) = \{\Delta_\sigma : x \in \Delta_\sigma\}$.

5.5.18. Lemma (Existence Lemma for $E\phi$, K_i, $\Diamond_i^\leq$ and $\Diamond_i^<$.). *If $\Diamond_i\phi \in \Delta_\sigma$ then there is a $\Delta_{\sigma'} \in W$ such that $\phi \in \Delta_{\sigma'}$ and $\Delta_\sigma \sim_i^\Gamma \Delta_{\sigma'}$. Similarly for $\Diamond_i^\leq$, $\Diamond_i^<$ and $E\phi$. Furthermore, if $\phi \in \Delta_\sigma$ then for all Δ'_σ, $E\phi \in \Delta'_\sigma$.*

Proof. Blackburn et al. [2001, p.442] for K_i and the preference modalities. The argument for $E\phi$, including the "furthermore" part, is a direct application of Lemma 5.5.13. ■

5.5.19. Lemma (Existence Lemma for I_i). *If $I_i\phi \in \Delta_\sigma$ then $|\phi| \in \iota_i^\Gamma(\Delta_\sigma)$.*

Proof. Trivially follows from the definition of ι_i^Γ. ■

5.5.20. Lemma (Truth Lemma). *For all $\phi \in \Gamma$, $\mathbb{M}^\Gamma, \Delta_\sigma \models \phi$ iff $\phi \in \Delta_\sigma$.*

Proof. As usual, by induction on ϕ. The basic cases, including the nominals, are obvious. Now for the inductive cases:

- $\phi = E\psi$, $\phi = K_i\psi$, $\phi = \Diamond^\leq\psi$ and $\phi = \Diamond^<\psi$. Standard modal logic argument from Lemma 5.5.18.
- $\phi = I_i\psi$. [Pacuit, 2007, p.26], from Lemma 5.5.19.

■

All that remains to show is that $\mathbb{M}^\Gamma$ is indeed a game model. We start by looking at the epistemic relation $\sim_i^\Gamma$.

5.5.21. Lemma (Adequacy of $\sim_i^\Gamma$ - Part I). *The relation $\sim_i^\Gamma$ is an equivalence relation.*

Proof. All $S5$ axioms are canonical [Blackburn et al., 2001, p.203]. ■

This means that $\{[\Delta_\sigma]_i : \Delta_\sigma \in W^\Gamma\}$ partitions the set W^Γ, for each agent. We can look at these partitions directly as strategies. That is, for each "profile" Δ_σ, set $\Delta_\sigma(i) = [\Delta'_\sigma]_i$ such that $\Delta_\sigma \in [\Delta'_\sigma]_i$. By the previous lemma we automatically get that this function is well-defined. The rest of the adequacy lemma for $\sim_i^\Gamma$ is then easy.

5.5.22. Lemma (Adequacy of $\sim_i^\Gamma$ - Part II). *For all Δ_σ and Δ'_σ, if $\Delta_\sigma \sim_i^\Gamma \Delta'_\sigma$ then $\Delta_\sigma(i) = \Delta_{\sigma'}(i)$ and $\iota_i^\Gamma(\Delta_\sigma) = \iota_i^\Gamma(\Delta'_\sigma)$.*

Proof. The first part is a trivial consequence of the way I set up $\Delta_\sigma(i)$. For the second part, observe that by the definition of ι_i^Γ all we need to show is that for all $|\phi| \in \iota_i^\Gamma(\Delta_\sigma)$, $|\phi|$ is also in $\iota_i^\Gamma(\Delta_{\sigma'})$. So assume the first. This means that $I_i\phi \in \Delta_\sigma$, which means by (K-I) that $K_iI_i\phi$ is also in Δ_σ. But then, because $\Delta_\sigma \sim_i^\Gamma \Delta_{\sigma'}$, we obtain by a routine modal logic argument that $I_i\phi \in \Delta_{\sigma'}$, which is just to say, $|\phi|$ is also in $\iota_i^\Gamma(\Delta_{\sigma'})$. ■

5.5.23. Lemma (Adequacy of $\succeq_i^\Gamma$ and $\succ_i^\Gamma$). *The relation $\succeq_i^\Gamma$ is a total, reflexive and transitive relation on W^Γ, and $\succ_i^\Gamma$ is its irreflexive and transitive sub-relation.*

Proof. The $S4$ axioms for $\Diamond_i^\leq$ are canonical. Irreflexivity of $\succ_i$ and totality of $\succeq_i$ are respectively enforced by the pure axiom (Tot) and (Irr), which are also canonical [ten Cate, 2005, p.69]. (Inc$_1$) finally ensures that $\succ_i$ is indeed a sub-realtion of $\succeq_i$. ■

5.5.24. Lemma (Adequacy of ι_i^Γ). *For all Δ_σ, $\iota_i^\Gamma(\Delta_\sigma)$ is a filter and does not contains the empty set.*

Proof. The filter part follows directly from K for I_i. See [Pacuit, 2007, p.29]. The second part is follows from (Ser). ■

5.5.5 Complete axiom system for $D\mathcal{L}_{\mathcal{GF}}$

I define $\Lambda_{D\mathcal{L}_{\mathcal{GF}}}$ as $\Lambda_{\mathcal{L}_{\mathcal{GF}}}$ together with the formulas in Table 5.3. Showing completeness boils down to show soundness for these new axioms.

5.5.25. Theorem (Soundness). *The formulas in Table 5.3 are sound with respect to the class of models for epistemic game frames and the restriction operation defined in 5.3.2.*

Proof. Soundness of the first four axioms is well known. It remains to show soundness of the fifth.

Take an arbitrary pointed model for game a structure $\mathbb{M}, w$, and assume that $\mathbb{M}, w \models \phi$ (otherwise we are done) and that $\mathbb{M}, w \models [\phi!]I_i\psi$. This means that $||\psi||_{|\phi} \in \iota_{i|\phi}(w)$, with $||\psi||_{|\phi} = \{w' \in W_{|\phi} : \mathbb{M}_{|\phi}, w' \models \psi\}$. Now, this can happen in only two cases.

1. $||\phi|| \cap \downarrow\iota_i(w) \neq \emptyset$ and $||\psi||_{|\phi} \in \uparrow(||\phi|| \cap \downarrow\iota_i(w))$. Now, unpacking the definition of $\mathbb{M}, w \models \mathsf{i}_i\phi$ reveals that this happens iff there is a $w' \in W$ such that for all $X \in \iota(w)$, $w' \in X$ and $w' \in ||\phi||$. But this is just to say that $||\phi|| \cap \downarrow\iota_i(w) \neq \emptyset$. Now that $||\psi||_{|\phi} \in \uparrow(||\phi|| \cap \downarrow\iota_i(w))$ means that $(||\phi|| \cap \downarrow\iota_i(w)) \subseteq ||\psi||_{|\phi}$. But this is the same as to say that $\downarrow\iota_i(w) \subseteq \neg||\phi|| \cup ||\psi||_{|\phi}$, where $\neg||\phi||$ is the complement of $||\phi||$ with respect to W. But $\neg||\phi|| \cup ||\psi||_{|\phi}$ is the same set as $||\phi \to [\phi!]\psi||$, which means that $||\phi \to [\phi!]\psi|| \in \uparrow\downarrow\iota_i(w)$. Since this last set is nothing but $\iota_i(w)$, this means that $\mathbb{M}, w \models I_i(\phi \to [\phi!]\psi)$.

2. $||\phi|| \cap \downarrow\iota_i(w) = \emptyset$ and $||\psi|| \in \iota_{|\phi}(w)$. In this case I defined $\iota_{|\phi}(w)$ as $\{W_{|\phi}\}$. This means that the second clause boils down to $||\psi||_{|\phi} = W_{|\phi}$. But this happens iff for all $w' \in \mathbb{M}_{|\phi}$, $\mathbb{M}_{|\phi}, w' \models \psi$, which in turn is nothing but to say that $\mathbb{M}_{|\phi}, w \models A\psi$. Assuming that $\mathbb{M}, w \models \phi$, this is the same things as $\mathbb{M}, w \models [!\phi]A\psi$. Now, for the first clause, observe that $\mathbb{M}, w \models \neg\mathsf{i}_i\phi$ is just the same as $\mathbb{M}, w \models I_i\neg\phi$. This means that $||\neg\phi|| \in \iota_i(w)$, which is the same as to say that $\neg||\phi|| \in \iota_i(w)$, i.e. $\downarrow\iota_i(w) \subseteq \neg||\phi||$, which happens iff (recall that $\iota_i(w)$ is a filter) $||\phi|| \cap \downarrow\iota_i(w) = \emptyset$.

■

5.5.6 Dynamic characterization of Nash equilibrium.

The crucial condition in the characterization of Section 5.2.2 is the mutual knowledge of each other's action. This, in turn, is the sort of condition that can typically be achieved by public announcements. In that case the very announcement that the agents play such-and-such a strategy surely does the trick.

5.5.26. Fact. [Nash equilibrium in $D\mathcal{L}_{\mathcal{GF}}$] Given a game model $\mathbb{M}$ with two agents, if at a profile w named by σ,

$$\mathbb{M}, w \models [\sigma(2)!]WR_1 \wedge [\sigma(1)!]WR_2$$

then w is a Nash equilibrium.

Proof. The argument again boils down to showing that at w both agents play a best response. Consider player 1. Clearly $\mathbb{M}, w \models \sigma(2)$, and so it must be that $\mathbb{M}_{|\sigma(2)}, w \models WR_1$. Now observe that $W_{|\sigma(2)} = \{w' : w' = w[s/w(i)]$ for a $s \in S_i\}$. But this means, by the same argument as in Fact 5.2.12, that $[w]_{i|\sigma(2)} = \{w\}$. So $\mathbb{M}_{|\sigma(2)}, w \models WR_1$ boils down to say that for all $s \in S_i$ and $w' = w[s/w(i)]$, $w' \succeq_i w$, as required. The argument for player 2 is symmetric. ■

We thus have a third characterization of the Nash equilibria, where they are now described as those profiles where the choice of a player would still remain rational after learning the other player's actions.

Chapter 6

Hybrid pragmatism, acceptances and norms on intentions

In the foregoing chapters I have explored how the volitive and the reasoning-centered commitment of intentions influence the deliberations of planning agents. As I mentioned in the Introduction, Section 1.2, intentions exert this influence because they are subject to four rationality constraints. They must be internally consistent, consistent with the beliefs, agglomerative and means-end coherent.

In this chapter I leave the formal theory and try to understand where these normative requirements come from. That is, I examine various ways to account for the fact that intentions are subject to these rationality constraints.

This issue has attracted much attention in recent years. Some authors have argued that the norms on intentions—which are usually thought of as practical norms—stem in fact from analogous norms of theoretical rationality associated with beliefs. This is the *cognitivist* view, championed for example by Harman [1976, 1986], Velleman [2003, 2005] and Setiya [2007, forthcoming]. Others, chiefly Bratman [2006b], have rather proposed an "agency approach", which avoids this reduction of the practical to the theoretical by accounting for the norms on intentions solely in pragmatic terms.

Both sides have their pros and cons. But, with the notable exception of Wallace [2006, 2003a], very few authors have moved away from these two extremes. There is, however, sufficient conceptual space for manoeuvre between pure cognitivism and the agency approach. In this chapter I investigate how far one can get in accounting for the normative requirements on intentions by using such an "intermediate" approach. This approach, based on the concept of *acceptances in deliberation*, tries to derive the norms on intentions from similar norms on acceptances. As we shall see, this is an essentially mixed approach because acceptances are "hybrid" attitudes, definitely on the cognitive side but still responsive to practical concerns. For that reason, I call it *hybrid pragmatism.* I argue that it provides a reasonable compromise between cognitivism and the agency approach, in a way that does justice to both the practical and theoretical aspects of the norms on intentions.

In Section 6.1 I return in greater detail to the four normative requirements on intentions that I presented in the Introduction. Section 6.2 is devoted to the cognitivist derivation of these requirements. The key sections of this chapter are 6.3 and 6.4, in which I introduce acceptances in deliberation and study how they can account for the norms on intentions.

In contrast to the preceding chapters, the investigation here is not formally driven. My goal is to provide the theory that I have developed so far with a philosophically solid basis. It will come clear as we move along, though, that hybrid pragmatism, with its focus on acceptances in deliberation, also introduces new issues onto the agenda for more formal enquiry. As such it helps us to understand better the functions of intentions in practical reasoning while opening up further research directions.

6.1 Constraints on rational intentions

In the Introduction (Section 1.2) I mentioned that *rational* intentions are required to be internally consistent, strongly belief-consistent, agglomerative and means-end coherent. Here I present these norms in more detail.

Means-end Coherence. A plan that aims to achieve some ends must contain intentions about necessary means[1]. More precisely, the agent must intend to do what he believes is necessary for him to intend to reach his end. Or, at least, he must plan to form the appropriate means-intentions later.

It is crucial that the "necessary means" are those that the agent has to *intend* to reach his end. Take for example an agent who intends to bring his oven up to a certain temperature, say to bake a cake. He might also know that this boils down to transforming electrical energy into heat. But he does not seem to be means-end incoherent if he does not have *per se* the intention to transform energy. To be sure, he has to intend to turn on the oven, which might require other actions such as turning some switches, but even though transforming energy is from a certain point of view a necessary means to achieve the required intention, it is not one that has to be directly intended for it to happen. The proviso on means-end coherence is precisely intended to cope with such cases.

The cases of means-end incoherence that I discuss feature a "gap" in the intention structure of the agent. That is, these are cases such that the agent intends an end E, believes that to achieve E he must form an intention about some means M, but does not have that intention. Of course, one can also conceive of stronger cases of means-end incoherent plans. An agent can intend an end E,

[1] Many authors have discussed the interpretation of this principle. The reader can consult Harman [1976], Brunero [forthcoming], Wallace [2006], Kolodny [2007], and Setiya [forthcoming].

believe that to achieve E he must come to form intentions about some means M but, instead, forms an intention that excludes his doing M. Such cases can in general be reduced to violations of one of the other requirements, and so I do not treat them as violations of means-end coherence.

Strong Belief Consistency. Means-end coherence crucially involves what the agent believes. But this connection between intentions and beliefs goes further. An intention should be feasible in a world where the agent's beliefs are true. That is, a plan should not be impossible to realize, given what the agent believes. This is what Bratman [1987, p.31] calls the *strong belief consistency* requirement[2].

Internal Consistency. Plans and intentions themselves have to be consistent. First, the content of an intention should be consistent. An agent should not intend impossible things, for example to do A and not to do A. Let me call this the internal consistency *of intentions*. But it also seems reasonable to ask plans to be consistent as wholes. The intentions in a plan should not contradict each other, they should not preclude one another's achievement. This can be called internal consistency of *plans*. Observe that a plan can be internally inconsistent even if each of its intentions is internally consistent. Internal inconsistency of plans arises out of the *combination* of the intentions it contains, a phenomenon that naturally brings us to the next requirement.

Agglomerativity. The original formulation of the principle of agglomerativity goes as follows:

> Given the role of intentions in coordination, there is rational pressure for the agent to *put his intentions together* into a larger intention. [Bratman, 1987, 134, my emphasis]

As I mentioned in the Introduction (Section 1.2), one can distinguish two readings of this principle[3]. First, there is what I called *agglomerativity as closure*: an agent cannot rationally intend to A and intend to B unless he also intends both A and B. What Bratman thus meant by "put intentions together" is, according to this interpretation, that they should close under conjunction.

For Yaffe [2004] this principle demands too much effort from agents with limited time and capacities. "It demands mental labor from agents that they have no need to perform, given their aims." In other words, agglomerativity as closure requires worthless combinations of intentions. He illustrates his claim

[2] Note that an agent does not have to believe that his plan is impossible to realize for it to be belief inconsistent. This is a stronger condition, to which I shall return soon. In Sections 6.2.4 and 6.4.4, however, I directly adopt this stronger condition.

[3] The two interpretations of agglomerativity, and the quotations in the following paragraphs, come from Yaffe [2004, p.511-512].

with the following example. Suppose one has the intention to go to Los Angeles tomorrow and the intention to go to London a year from now. To him, it is not worth the effort to combine these two intentions into a single one. To have the merged intention neither contributes to the achievement of the individual ones nor does it help the agent to coordinate his own actions. So, according to Yaffe, agglomerativity as closure demands that one combines intentions that do not need to be put together. Since the combination itself takes a certain amount of time and effort, it should not be required of normal, i.e. resource-bounded, agents.

In view of this he proposes the following alternative interpretation of the principle, which I called *agglomerativity against potential irrationality*: it is irrational to intend A and to intend B if the intention to do both would be irrational according to some other norms of rationality for intentions. According to this second interpretation planning agents are no longer required to combine arbitrary intentions. Rather, they are required to do so only to the extent that this "makes conflicts evident to themselves, when there is doubt as to the rationality of the conjunctive intentions". The conflicts mentioned here are conflicts "with other norms of rational intentions". Given what I have said so far, this means that having two intentions is irrational if their combination would result in an intention whose content is contradictory or impossible to realize given the agent's beliefs[4].

Observe that, understood that way, agglomerativity parasitizes, so to speak, its rationality demands on these two other norms. In the case of internal consistency, for example, having the intention to do A and the intention to do B is irrational to the extent that having the intention to do A and B is internally inconsistent. Along the same lines, having the intention to do A and the intention to do B, given the belief that A and B cannot be achieved together, is irrational to the extent that having the intention to do A and B is strongly belief-inconsistent.

It is thus no coincidence that the former case resembles what I have called above "internal inconsistency of plans". Plans were called internally inconsistent precisely when a pair of their elements could not be achieved together[5]. In other words, internal consistency of plans is a particular case of agglomerativity against potential irrationality, preventing the agent from holding pairs of intentions which would violate internal consistency of intentions if they were put together.

It also worth noting that, unlike agglomerativity as closure, agglomerativity against potential irrationality does not require a systematic combination of inten-

[4] Recall that I consider means-end incoherent plans as plans where the means-intentions are missing. But agglomerativity as potential irrationality is about pairs of intentions the agent already has. In view of that, a violation of that requirement that resorts on means-end coherence seems to involve what I have called strongly means-end incoherent plans. But, as I said, it seems that these cases are, in turn, reducible to violation of some other norms. For that reason I only consider cases where agglomerativity as potential irrationality leads to a violation of internal or strong belief consistency.

[5] It is, in that respect, telling that Yaffe does not consider internal consistency of plans as a separate requirement.

tions. Combining individual intentions into a single intention with a conjunctive content somehow becomes instrumental to the unveiling of other forms of irrationality. As long as the intentions of an agent are not potentially problematic, the agent is not required to agglomerate them.

In what follows I work with these two forms of agglomerativity in parallel, for I do not consider Yaffe's argument for rejecting agglomerativity as closure totally convincing, while I do not find agglomerativity against potential irrationality completely satisfactory either. Let me sketch the reasons for my doubts before going further.

Granted, it is unrealistic to ask limited agents to agglomerate all their intentions into a single "grand world" intention[6]. But it is rather unlikely that this is what the original idea of agglomerativity was intended to mean. Rather, it seems more probable that it was aimed at smaller "worlds", maybe at the level of plans, where agglomeration *should* be systematic. It seems plausible that putting the various intentions of a plan together does facilitate personal coordination in extensive decision problems, for example, even though this combination does not reveal any violation of the other rationality requirements. If this is so, then Yaffe's overall rejection of agglomerativity as closure goes too far. On the other hand, it also means that agglomerativity against potential irrationality does not tell the whole story. There may be cases where it is rational to agglomerate systematically.

Let me summarize. I started with four rationality requirements on intentions: means-end coherence, strong belief consistency, internal consistency and agglomerativity. Internal consistency applies within intentions and within plans. The last reguirement, though, is a special case of agglomerativity against potential irrationality, namely when there is a threat of violating internal consistency within intentions. Following Yaffe I distinguished this form of agglomerativity from agglomerativity as closure. So there are really five rationality requirements to explain: means-end coherence, strong belief consistency, internal consistency within intentions, agglomerativity as closure and agglomerativity against potential irrationality[7].

6.2 Cognitivism

In this section I present the *cognitivist* view, according to which the norms just presented ultimately stem from norms of theoretical rationality associated with beliefs.

[6] I draw this appellation from Savage [1954]. My reservations with respect to Yaffe's parallels the difficulty of explaining how agents set the "size" of their worlds, i.e. of their decision problems.

[7] There are of course interrelations between these various norms. See again Section 2.2.

The keystone of cognitivism is a postulated connection between intentions and beliefs. I present two ways to understand this connection. Most of this section is then devoted to seeing how well they support the derivation from the theoretical norms on beliefs to the practical norms on intentions. All of this, of course, rests on a particular philosophical theory of beliefs. In the previous chapters we could do most analyses without diving into such conceptual details. For the present chapter, though, they become crucial, especially in distinguishing between beliefs and acceptances in deliberation.

6.2.1 The functionalist view on beliefs

Many authors who have written on the relation between practical and theoretical rationality take good care to distinguish between *probabilistic* and *flat-out* [Bratman, 1987] or *all-or-nothing* [Harman, 1986] beliefs. They differ mainly in that the first, in contrast to the second, comes in various degrees. Probabilistic beliefs echo *subjective probabilities* or *endogenous uncertainty*, which I mentioned in the Introduction (Section 1.1.1). Flat-out beliefs, on the other hand, are attitudes that an agent either has or not. In that respect, they are close in behaviour to the qualitative representations of information that I used in most of the previous chapters[8].

Whether probabilistic or flat-out, the theory of beliefs that underlies most cognitivist accounts is *functionalist.* Just as intentions, beliefs are characterized through their "actual and potential, or typical, causal relations to sensory stimulations, behavior and other mental states" [Schwitzgebel, 2006].

A belief that p is thus viewed as *an attitude of regarding p as true* that[9]:

1. dispose the subject to incorporate p into further practical and theoretical reasoning.

2. is "formed, revised and extinguished—or [...] for short [...] *regulated* for truth", in the sense that it is responsive to evidence and reasoning.

3. is correct if and only if p is the case.

Once again following Schwitzgebel [2006], one can think of the first two conditions as respectively specifying the backward-looking and forward-looking conditions for an attitude to functionally count as a belief. I used this idea to describe

[8] It is, however, unclear whether flat-out beliefs match the subject matter of epistemic logic. For one thing, epistemic logic is often thought of as the study of sure beliefs, which correspond in turn to probabilistic beliefs of degree 1 or 0. But it is often stressed, e.g. by Bratman [1991], that flat-out beliefs are *not* reducible to beliefs with degree 0 or 1. I do not go into this distinction in detail, for it is orthogonal to my present concern. I simply proceed, unless explicitly stated, with the two notions of belief.

[9] This characterization is taken almost textually from Shah and Velleman [forthcoming, p.2-3].

intentions in the Introduction (Section 1.2). Intentions are *outputs* of practical reasoning (backward-looking) which typically also play a certain role as *inputs* to deliberation (forward looking).

Condition (2) is backward-looking in the sense that it points to processes that *affect* beliefs. They are created, revised or abandoned according to the input of new evidence or conclusions reached from theoretical reasoning. This is a *constitutive* claim about what beliefs are.

This should be contrasted with condition (3), which gives a criterion for correctness, and which is thus essentially *normative.* It states that beliefs are correct only to the extent that they are true, i.e. that they fit the world[10]. In the words of Austin [1953] and Searle [1983], beliefs have the "mind-to-world" direction of fit.

Conditions (2) and (3) sharply distinguish beliefs from *practical* attitudes like intentions. Intentions can be formed, revised and abandoned in response to changes in desires or preferences, and their correctness is more often than not a matter of instrumental rationality. Cases of beliefs formed or held for practical reasons are, on the other hand, not typical and are usually viewed as a form of wishful thinking. That is, these are pathological cases, i.e. incorrect ways to hold belief. For that reason, beliefs are often said to belong to the realm of *theoretical* rationality.

This is not to say, of course, that they do not take part in practical reasoning. Condition (1) makes this clear. It is "forward-looking" because it points to typical processes that take beliefs as *input.* Although there might be other forward-looking conditions that characterize beliefs, the disposition to be incorporated into practical and theoretical reasonings is the most important for our present concerns[11].

6.2.2 Constraints on rational beliefs

I have already introduced a normative component in the functionalist definition of beliefs, but the cognitivist derivation of the practical norms on intentions does not explicitly use it. Instead, it appeals to three other normative requirements on beliefs, which I now present.

[10] Here I follow Shah and Velleman [forthcoming] and include the normative claim in the definition of beliefs. This is by no means an uncontroversial practice, but it will prove useful in distinguishing beliefs from acceptances in deliberation.

[11] It is worth noting that one even finds characterization of beliefs only in terms of potential input into practical reasoning. Holton [1994, p.68], for example, says that " your belief that a certain thing will happen is just the disposition that you acquire when you work the supposition that it will happen into your plans." Similar remarks can be found in Schwitzgebel [2006] and Alonzo [forthcoming]. Note, furthermore, that this view of belief quite nicely matches the Bayesian approach to belief that is inherent in representation results in decision theory. See the references in the Introduction (Section 1.1.1) and Joyce [2004].

Internal Consistency of Beliefs. Just as with intentions, one can distinguish two senses of this requirement. Internal consistency *within* beliefs requires the content of beliefs to be consistent. An agent should not believe contradictions. Internal consistency *between* beliefs, on the other hand, asks the different beliefs of an agent to be consistent with each other. As was the case with intentions, internal consistency between beliefs follows from internal consistency within beliefs, together with agglomerativity, which I shall introduce shortly. For that reason, I use the plain "internal consistency of beliefs" to refer to internal consistency within beliefs.

Agglomerativity of Beliefs. Again, one can think of agglomerativity as closure under conjunction or as a safeguard against potential violations of other norms on beliefs. In the case of intentions, this second interpretation of the principle was proposed as an alternative to the first one, which Yaffe [2004] found too strong. Agglomerativity as closure is less controversial for beliefs. Here is Velleman [2003, p.18] on the subject:

> Beliefs are agglomerative because they aim to fit the world, of which there is just one, in whose complete characterization the contents of all true beliefs are conjoined. The rational pressure to conjoin beliefs is a pressure to fuse them into a single characterization of the single world that all of them aim to fit.

As this last sentence suggests, there seems to be a relation between the standard of correctness for beliefs mentioned in Section 6.2.1 and the fact that they are agglomerative. Beliefs have to fit the world, and for that reason it seems that there is a rational requirement to agglomerate them.

One can obtain agglomerativity against potential irrationality of beliefs if they are closed under conjunction. This is so because, for beliefs, agglomerativity against potential irrationality really amounts to agglomerativity against potential *inconsistency*[12]. An agent whose beliefs are closed under conjunction and who has "potentially inconsistent" beliefs would turn them directly into an internally inconsistent conjoined belief. For that reason, I take "agglomerativity of beliefs" to mean only agglomerativity as closure.

Explanatory Coherence. The last requirement, which I call "explanatory coherence" [Harman, 1986], comes from what Schwitzgebel [2006, Section 3.2] calls the "holistic" view on beliefs. In normative terms, it requires a rational agent to maintain "relations of immediate coherence or intelligibility" among his beliefs [Harman, 1986, p.75][13]. That is, given one particular belief of an agent,

[12] In Section 6.1 I made a proviso regarding means-end coherence. The same applies to explanatory coherence, which I introduce next.

[13] Holism about beliefs also has a constitutive counterpart. See again Schwitzgebel [2006, Section 3.2] for more explanations and references.

one should be able to locate other beliefs that "explain" it[14]. Given that beliefs are responsive to evidence, this means that a belief that ϕ should be backed, so to speak, by a belief about *evidence* in favour of ϕ.

6.2.3 Weak, intermediate and strong cognitivism

There are thus three norms on rational beliefs: internal consistency, agglomerativity and explanatory coherence. For the cognitivists the requirements on intentions that I presented in Section 6.1 can be derived from these requirements on beliefs. This derivation typically rests on the assumption that intentions "involve" beliefs. One can classify the various cognitivist derivations according to their view on how strong this involvement is. More precisely, Bratman [2006b] distinguishes three strengths of cognitivism, which I now present. Even though only the last two will be of interest hereafter, a close look at the first one will help to clarify issues.

The requirement of strong belief consistency is already an assumption on how intentions involve beliefs, but it is a negative involvement, so to speak. It states that having the intention to do A implies not believing that one will not do A. *Weak* cognitivism holds something slightly stronger: if an agent intends to do A he must consider it possible that he will do A[15]. In other words, to have the intention to do A implies believing that it is possible that one will do A. This belief can be either flat-out, as advocated e.g. by Wallace [2003a, 2006], or probabilistic, as Chan [1999] takes it. In both cases, however, this belief is compatible with the belief that it is possible that one will *not* do A, and both can be *simultaneously and consistently* incorporated into practical reasoning as partial beliefs. That is, the agent can work into his further deliberation the fact that he might and that he might not do A.

I call the second type of cognitivism *intermediate*. It holds that having the intention to do A implies believing that one will do A. This, of course, entails weak cognitivism, but not the other way around. As I take it, the key idea underlying intermediate cognitivism is that intending to do A implies, first and foremost, being disposed to work in *flat-out* the fact that one will do A in further planing and, second, that this assumption is regulated by truth and is responsive to evidence. That is, it seems that intermediate cognitivism is not compatible with incorporating the fact that one might not do A in practical reasoning, once one intends to do A. Or at least this is what I take intermediate cognitivism to mean: intending to do A implies believing that A in the sense of being disposed

[14] I leave open the question of what should be necessary or sufficient to count as an explanation here. See the illuminating discussion of Harman [1986] and the papers cited in the footnote on page 138. I return briefly to this issue in Section 6.3.

[15] This is stronger than strong belief consistency as long as we assume that not believing that not ϕ is not equivalent to considering it possible that ϕ. Recall that this equivalence holds, for knowledge, in all the epistemic models that I used in the last chapters.

to use the fact that one will do A in deliberation, in a way that is regulated by truth and is responsive to evidence.

Except for Harman [1986], very few authors have directly argued for the intermediate cognitivist view. The most popular cognitivist standpoint is rather stronger. According to e.g. Harman [1976], Velleman [2005] and Setiya [2007], intending to do A *is the same as* having a special kind of belief that one will do A. I call this *strong* cognitivism. Again, the kind of belief involved here seems to be essentially flat-out, at least as far as integration into practical reasoning is concerned. Strong cognitivism thus holds that to have the intention to do A is nothing else than being disposed to work in the fact that you will do A in your practical reasoning, in a way that is responsive to evidence and regulated by truth[16].

6.2.4 Rational intentions from rational beliefs

I now show how cognitivists derive the norms on intentions from the norms on beliefs. I look at each of the derivations in some detail, for they will serve as a landmark for the "hybrid" derivation that I present later.

Internal and strong belief consistency of intentions. These two norms are the easiest to derive for the cognitivist. Let me look first at internal consistency of intentions. Suppose one intends something contradictory, say to do A and not to do A. By the intermediate cognitivist assumption, one then must believe that he will do A and that he will not do A, which violates internal consistency of beliefs. Putting back this contrapositive argument in its normal direction, it shows that internal consistency of beliefs and the intermediate cognitivist assumption together imply internal consistency of intentions.

The argument for strong belief consistency goes along the same lines. An intention that is not feasible given the agent's background beliefs will generate a new belief that is, by assumption, inconsistent with this background. But then, using agglomerativity, one gets a belief that is internally inconsistent[17].

Means-end from explanatory coherence. To derive means-end coherence of intentions from explanatory coherence of beliefs, cognitivists usually make one more assumption about what counts as evidence for an agent that he will act in a certain way. As I have mentioned many times now, beliefs are taken to be responsive to evidence. But in the case of beliefs about what one will do, it seems that we have a special source of evidence[18], namely the intentions themselves.

[16] The authors mentioned above usually add special conditions to the way these intention-as-beliefs are responsible to evidence and regulated by truth. I leave these details aside.

[17] Observe that one could have argued directly from internal consistency between beliefs.

[18] I am using "evidence" rather sloppily here. For in-depth discussions about the epistemology of agency, see Anscombe [1957], Faley [2000] and Velleman [2005].

This means that such beliefs have to be explained by beliefs about what the agent intends.

Cognitivists derive means-end coherence of intentions from explanatory coherence of beliefs as follows. Suppose that one has means-end incoherent intentions. He intends the end E but he does not intend any of what he believes are means M that he must come to intend to achieve E. What is more, he does not intend to form such means-intentions later. By the intermediate cognitivist assumption, this means that this agent must believe that he will do E. But since he also believes that he will do E only if he does something in M, he should believe that he will do some M[19]. But since he does not intend to do anything in M, and he does not believe that he has such intentions, it seems that he lacks the required evidences to back up this new belief, so to speak. In other words, the beliefs of the agent are explanatory incoherent. It thus seems that explanatory coherence, together with the fact that beliefs about intentions count as evidence for beliefs about what an agent will do, imply means-end coherence.

This derivation is unfortunately not sound. It slips too quickly from one not having the required means-intentions to one not believing that one has these intentions. As pointed out by Brunero [forthcoming], Bratman [2006b] and Wallace [2006, 2003a]), the latter can come without the former. Here is Bratman [*idem*]:

> Suppose I intend E and know that E requires both M and that I intend M. If I still do not intend M my intentions suffer from means-end incoherence. But suppose that, while I in fact do not intend M, I nevertheless falsely believe that I intend M. So my beliefs are that E, that E requires both M and that I intend M, that I intend M, and that M. There is no incoherence (though there is falsity) in this structure of beliefs. So means-end coherence is not belief coherence.

Bratman points out that an agent can falsely come to believe that he intends something and so he has coherent beliefs, but nevertheless incoherent intentions. To carry the above derivation through, cognitivists thus need not only to assume that the intentions are means-end coherent, but also that the agent is not mistaken about what he intends.

This, according to Wallace [2003a, p.21], cannot be assumed without some additional rationality constraints: "theoretical constraints on rational beliefs can get you as far as the *belief* that you intend to do [something]; to go beyond that, to a rational requirement that you form the [required means-intentions], we need an additional principle [...]", a principle that is independent of explanatory coherence of beliefs.

The principle that Wallace has in mind is a pragmatic one. He holds that to have true beliefs about one's own intentions is, "*in deliberative contexts where [means-end coherence] is relevant*, [...] an executive virtue, to be included among

[19] This step assumes, of course, something like an inference closure principle for beliefs.

the traits and capacities that make us, in general, effective in the pursuit of our goals" or "is a strategy that enhances our ability to realize the broader aims that are given with our nature as deliberating agents" [Wallace, 2006, p.119, my emphasis]. That is, having a false belief about one's own intentions, when that belief features in the cognitive background of deliberation, seems to threaten the overall prospect of reaching one's ends. Bluntly, being wrong about what you intend is, in the long run, not good.

This is an important shift in Wallace's argument, because he thereby steps outside the pure cognitivist enterprise. In his view, it is not "plausible to suppose that the whole field of instrumental rationality can be explained in terms of the requirements of coherence in beliefs" [*idem*]. This recourse to the principle of practical rationality will also be very important later on. I argue in Section 6.4 that this is precisely the kind of hybrid justification that comes out of using acceptances instead of beliefs to derive the norms on intentions.

This is looking too far ahead, however: for now what matters is to observe that Bratman [2006b, p.13] has offered a counterexample to the thesis that mistaken beliefs about one's own intentions are irrational. He pointed out that "to reflect carefully on all that one intends [...] is an activity that takes time and uses other resources, and one may well have better things to do." Bratman considers, for example, an agent who believes that he intends to go shopping on Thursday while, in fact, he intends to go on Friday. A week before, it seemed perfectly rational for this agent not to reflect on the accuracy of this belief because "other matters are more pressing right now"[*idem*].

Bratman's example is rightly aimed at showing that it is not always irrational, practically speaking, to have false beliefs about one's intentions. But Wallace seems to have two ways around it.

First, he crucially uses the notion of relevant deliberations. It is notable that, in Bratman's counter-example, the agent is correct about the fact that he intends to go shopping, but he is mistaken as to *when* he intends to go. Now, the essence of the case is that it does not seem irrational to have an incorrect belief as long as its accuracy is not relevant to further deliberations. To see this, suppose that the agent intends to go shopping in order to buy some camping gear that he needs for a hiking trip the week after. The fact that he is mistaken about when he actually intends to buy the gear does not threaten his hiking plans. On the other hand, this mistaken belief becomes more problematic if he is to plan his Thursday evening. Whether or not it is irrational to have such mistaken beliefs depends on the context of deliberation, and this is precisely what Wallace seems to point out when he focuses on "deliberative contexts where [means-end coherence] is relevant [...]".

Second, Wallace also holds that being mistaken about what one intends is irrational because it threatens the *overall* prospect of achieving one's own intentions. In other words, not all instances of such mistaken beliefs have to mess up the agent. Rather, such beliefs are irrational because they embody, so to speak, a bad deliberation habit. In this context, is not so devastating that there exist cases of false beliefs about one's own intentions which are not irrational.

These two replies to what we may call the "problem of false beliefs" will also be very important later, because they apply even more naturally with acceptances in deliberation. But for now the reader should bear in mind that cognitivism has difficulty in accounting for means-end coherence. There seems to be a way to derive this norm on intentions from explanatory coherence of beliefs, if one is willing to use an additional pragmatic principle. But, as we shall see presently, means-end coherence is not the only norm that resists a purely cognitivist derivation.

Agglomerativity. Let us look first at agglomerativity against potential irrationality. As this requirement imports part of his rationality demands from other norms on intentions, it should not come as a surprise that cognitivism can explain it to the extent that it can explain the others. There are, more precisely, two cases to consider. Suppose first that an agent has the intention to do A and the intention to do B and that the combination of these into the intention to do A and B would be internally inconsistent. Clearly, this pair of intentions would be irrational because the combined intention would generate an internally inconsistent belief, given the intermediate cognitivist assumption[20]. An entirely similar argument covers the case where the combination of A and B would generate a strongly belief-inconsistent intention. So cognitivism can easily explain agglomerativity as potential irrationality, simply because it can explain internal and strong belief consistency.

The case of agglomerativity as closure is more problematic. Strong cognitivism can of course explain it, given agglomerativity of beliefs. After all, to hold that beliefs are agglomerative is just the same as to hold that intentions are also agglomerative, once one views intentions as a special kind of beliefs.

But observe that the argument does not go so straightforwardly for intermediate cognitivism, even if one assumes agglomerativity of beliefs. Suppose an agent has the intention to do A and the intention to do B, but not the intention to do A and B. From the assumption underlying intermediate cognitivism, together with belief agglomerativity, we can conclude that the agent believes that he will do A and B. But what is supposed to be wrong with having this belief while not having the corresponding intention? Note that he does not have the intention not to do A and B, which would simply make his beliefs internally inconsistent.

One possible way out would be to use explanatory coherence again. One would thus say that the belief that one will do A and B wants an explanation, which the belief that one has the intention to do A and B would provide. But why can the beliefs that one intends to do A and the belief that one intends to do B not together provide the required explanation? If they do, then we are still missing

[20] Note that one could also argue directly from these two intentions, the intermediate cognitivist assumption and agglomerativity of beliefs.

a intermediate cognitivist argument for agglomerativity.

What is more, as we have seen for means-end coherence, the recourse to explanatory coherence of beliefs makes one vulnerable to the problem of false beliefs about intentions. In the present context, the problem can be rephrased as follows. If an agent can be mistaken about his own intentions, i.e., can have false beliefs about what he intends, then his belief that he will do A and B can be explained by the (false) belief that he intends to do both.

In view of all this, it seems that intermediate cognitivism can explain internal consistency, belief consistency and agglomerativity as potential irrationality. It has difficulties with means-end coherence, because of the problem of false beliefs about one's own intentions. Strong cognitivism, however, explains agglomerativity as closure.

6.2.5 A general concern against cognitivism

The shortcomings of the cognitivist derivations of agglomerativity and means-end coherence are in themselves enough to motivate the search for an alternative. But Bratman [1987] has famously cast doubts on the basic assumption that intending to do A implies believing that one will do A. According to him, this assumption rules out some plausible cases of agents who apparently have the intention to do A while not believing that they will do A. He explains:

> [In [Bratman, 1987]], it seemed to me plausible that I might, for example, intend to stop at the bookstore on the way home even though I know that, once I get on my bike I tend to go on automatic pilot, and so even though I do not, strictly speaking, believe that I will stop (though I do not believe that I will not stop). So I thought it best not to tie the theory of intention and planning to such a strong belief condition. [Bratman, 2006b, p.3]

The "strong belief condition" that Bratman writes about here is what I have called intermediate cognitivism. Given what I said in Section 6.2.3, what is at stake is whether the agent believes *flat-out* that he will stop at the bookstore. For a cognitivist, if this agent really does not have such a flat-out belief then we cannot say that he has the intention to stop at the bookstore. Bratman, on the other hand, thinks that the agent can genuinely intend to do so even if he does not have this flat-out belief.

Recall that an agent counts as having such a belief if he regards the fact that he will do A as true in a way that:

1. disposes him to incorporate the fact that he will do A into further practical and theoretical reasoning.

2. is regulated by evidence in favour of him going to do A.

3. is correct if and only if he will in fact do A.

It seems to me that the only thing to deny is the fact that the agent, from having the intention to do A, is automatically disposed to work the assumption that he will do A into his *theoretical* reasoning. To be sure, one cannot deny that the agent regards the fact that he will do A as "true". This agent is disposed to incorporate this fact into *practical* reasoning[21]. What is more, as we have seen in Section 6.2.4, having the intention to do A can be seen as an evidence for the fact that one will do A, especially in the context where the intention at hand is future-directed. There is not yet a "fact of the matter" that can settle the correctness of this attitude. Given all this, the only thing left to deny is the disposition to incorporate "I will do A" into theoretical reasoning. This reading is supported, I think, by the original formulation of the absent-minded cyclist example:

> I might intend now to stop at the bookstore on the way home while knowing of my tendency towards absentmindedness—especially once I get on my bike and go into "automatic pilot." *If I were to reflect on the matter* I would be agnostic about my stopping there, for I know I may well forget. It is not that I believe that I will not stop; I just do not believe I will. [Bratman, 1987, p.37, my emphasis]

The reflection that is mentioned here seems to be essentially truth-oriented, in the sense that what is at stake is whether the agent will, in fact, stop at the bookstore. In other words, the agent deliberates about the truth of "I will stop at the bookstore" and not about, for example, whether he should stop or how he would make it happen.

In view of all this, I will take the general concern about cognitivism to be the following. Intending to do A seems to come with an attitude of regarding "I will do A" as true, but this attitude is not quite a belief. In particular, it does not dispose the agent to incorporate this fact into theoretical reasoning. The overall aim of this chapter is indeed to see whether the normative requirements on intentions can be derived if one thinks of this attitude not as a belief but rather as an acceptance in deliberation[22].

[21] Witness the discussion in Bratman [1987, p.37-39].

[22] It is interesting to note, before going further, that Bratman [2006b] does not take this route. As noted in the introduction of this chapter, he proposes a justification of the norms on intention that altogether bypasses the recourse to cognitive-like attitudes, whether beliefs or acceptances. For him norms on intentions stem from the general "aim to achieve what is intended" and of the "projected unity of agency". Thus the appellation "agency approach" to describe Bratman's approach.

6.3 Acceptances in deliberation

In this section I first introduce the concept of "acceptance in deliberations", which bases the alternative derivation of the norms of intentions. After that I briefly present the companion idea of "adjusted background of deliberation", which will be important hereafter, when I look at the normative requirements that apply to acceptances.

6.3.1 Accepting or taking for granted in deliberation

In the context of practical reasoning, some authors[23] observed that there is more than beliefs in the "cognitive background of deliberation". We sometimes intentionally *accept* or *take for granted* some facts about the world, even though we neither believe them with degree 1 nor flat-out. These phenomena are called "acceptances in a context". In what follows I am specifically concerned with the role of these attitudes in practical reasoning. For that reason I refer to them as acceptances *in deliberation*, or simply as acceptances. I should mention that Shah and Velleman [forthcoming] talk about acceptances as the general attitude of regarding something as true. For them, beliefs, assumptions and even images are all specific kinds of acceptances. As we shall soon see, the acceptances that I am concerned with, acceptances in deliberation, are a specific kind in this general category of epistemic attitudes.

Acceptances in deliberations are not only regulated by truth and responsive to evidence, but also also "regulated for practice" and responsive to "pragmatic considerations" [Alonzo, forthcoming]. This difference is best illustrated by an example.

> In planning my day—a June day in Palo Alto—I simply take it for granted that it will not rain even though I am not certain about this. If I were instead figuring out at what odds I would accept a monetary bet from you on the weather I would not simply take it for granted that it will not rain. But in my present circumstances taking this for granted simplifies my planning in a way that is useful, given my limited resources for reasoning. [Bratman, 1991, p.22]

Even though the agent, in this case, incorporates the fact that it will not rain into his practical reasoning, he does it in a peculiar way. Observe first that he

[23] Chiefly Bratman [1991]. For a congenial but nevertheless different characterization of acceptances, see Grice [1971], [Cohen, 1989] and [Engel, 1998]. Holton [1994] has also studied related phenomena in the context of thrusting relations. Harman [1976, p.438] already spoke of "taking for granted" in relations with intentions, but the appellation "acceptances" seems to come from Willams [1973]. Note that some authors—e.g. Holton [*item*] and Alonzo [forthcoming]—use "reliance" instead of acceptance to talk about what appears to be the same attitudes. Engel [1998, p.148] finally remarks that one finds discussion of "pragmatic beliefs" already in Kant, in a way that is very close to what will be described as acceptances.

does not plan with the idea that the chances of rain are extremely low, even though one could argue that this is what he really believes. He plans with the plain assumption that it will not rain. Observe too that what really triggers the incorporation of this fact into the agent's planning is a *pragmatic concern.* It simplifies deliberation. These seem to be the key features of acceptances. They can be at variance with the agent's beliefs and are responsive to pragmatic considerations.

In what follows I shall thus take an acceptance that p in a given deliberation to be an attitude of regarding p as true that:

1. Incorporates p into that very deliberation.
2. Is regulated by either the truth of p or the pragmatic context of a given deliberation, in a way that is responsive to evidence or practical concerns.
3. If both p is not true and the agent is not better off by accepting p, then the acceptance that p is not correct.

Acceptances thus share with beliefs the part of the "forward-looking" (see Section 6.2.1) functional characterization. They are both states that are crucially incorporated into practical reasoning. But I do not take acceptances in deliberation as featuring typically in theoretical reasonings, even though some authors, e.g. Stalnaker [1984], have studied closely related phenomena in that context. I take them to be specifically tailored for practical contexts.

Acceptances and beliefs differ in the way they are regulated, i.e. formed, changed or discarded. Condition (2) states that acceptances can be regulated both by evidence and practical concerns. Similarly, condition (3) states that practical concerns also come into play to assess the correctness of acceptances. Here I deliberately leave unanswered the question whether there is a systematic relation between the evidential and practical inputs, for example whether one has priority on the other. For what follows, it will be enough to know that acceptances are responsive to both[24].

[24] I think that to provide sufficient conditions for correctness one would have to be more precise about this relation, especially to handle properly the relation between bracketed beliefs and acceptances (see below). For that reason (3) only specifies necessary conditions for correctness.

Such a more precise stance could go as follows. Following Alonzo [forthcoming], I find it plausible to think that the pragmatic considerations that justify acceptances are constrained by evidence. That is, it seems that truth somehow has precedence over practice in terms of the correctness of acceptances. One could propose to make this precedence more explicit, namely in *lexicographical* terms. Condition (3) would then state that an acceptance that p is ultimately correct if p is the case. But in case the truth of p is not (yet) a settled matter then, and only then, can practical considerations enter the assessment of whether accepting that p is correct. This will happen when p is about something in the future, as in the example above, and especially about future *actions.* As observed again by Alonzo [forthcoming], this could help to distinguish acceptances from simple wishful thinking. Acceptances should respond to evidence in a way that wishful thinking does not. If there is any way that the later can be justified, it seems that this will be solely by practical concerns.

Observe that, in the last example, it is the acceptance in itself that seems to be practically useful. The practical consequences of the fact that it will rain do not really influence how one assesses the acceptance. This is what condition (3) makes precise: the pragmatic considerations should bear on the acceptance itself, rather than on its content. Pragmatic arguments that justify accepting that p are arguments that show that the agent can be better off by incorporating p into his practical reasoning, and not necessarily that the agent would be better off if p were the case. These can go together, but they need not[25].

The simple fact that practical considerations regulate and take part of the standard of correctness for acceptances suffices to distinguish them from beliefs. But this difference also shows in the context-dependency of these two attitudes. Bratman [1991] has strongly emphasized that the rationality of acceptances depends in general on the context of deliberation. An agent may rationally take some facts for granted in some deliberations while not taking them for granted in others. This is not the case for beliefs, either probabilistic or flat-out. Whether a belief is justified depends on the evidences that support it, not on the specific deliberation in which it is applied. To appreciate this difference, look again at Bratman's example. While planning his day, it seems perfectly justified for the agent to take it for granted that it is not going to rain. But it is also justified for him to abandon this acceptance when he has to decide how much he would bet on that very fact, even though the matter remains unsettled in both cases. To put it differently, it would not be rational for the agent to change his belief about the weather from one deliberation to the other, unless he receives new information in the meantime. But the change in acceptance seems perfectly justified. This, [Bratman, 1999, chap.1] argues, shows that acceptances are not the same as beliefs.

It should be noted that an agent can *decide*, to a certain extent, which facts he takes for granted. Following Bratman [1991, p.29], we can distinguish two ways in which an agent can take a fact A for granted: *positing* and *bracketing.* The first occurs when the agent decides to take A for granted even though the agent is uncertain about it. This is the case I had in mind in most of the previous discussion. A gets temporarily promoted, so to speak, from mere possibility to "hard" fact. Bracketing occurs, on the other hand, when an agent believes that A but decides not to work this assumption into his deliberation. Unfortunately, Bratman is not very explicit about bracketing. It seems to concern mainly flat-out beliefs or those with degree 0 or 1. In these cases, one can imagine that the agent decides to plan as if the possibility of not-A was serious, even though he does not believe so. For flat-out beliefs, note that this crucially uses the fact that the agent

[25] Take for example a chess player who takes for granted that his opponent will play very well, without having any conclusive evidence for that. We can easily imagine that this acceptance makes the player better off. He will plan several moves ahead and play more cautiously, which might make him more likely to win. But observe that if, in fact, the opponent does play well, then the chances of winning for the player seem to be actually diminished.

can decide what he takes for granted. Flat-out beliefs were indeed characterized as disposing the agent to include their contents into practical reasoning. By bracketing the agent explicitly overrides this disposition.

6.3.2 The adjusted background of deliberation

At the beginning of this section I introduced acceptances by way of the *(adjusted) cognitive background of deliberation.* Bratman [1991, p.29] put forward this concept to stress that the information invoked in practical reasoning is different from the collection of the agent's beliefs, which he calls the *default* cognitive background. The cognitive background of deliberation is "adjusted" precisely because it contains acceptances resulting from positing or bracketing elements of the default background.

I remarked earlier that acceptances and beliefs have a similar forward-looking functional characterization, at least as far as practical reasoning is concerned. They are attitudes that feature in the adjusted cognitive background of deliberation. The adjusted cognitive background of deliberation thus differs from the "default" background in that it features contents of acceptances *and* of beliefs.

The remarks at the end of the previous section suggest that agents can somehow build the adjusted background of deliberation by positing and bracketing some facts. But agents with limited time and resources can only make a small number of such explicit decisions[26]. Should we take, then, the adjusted cognitive background to contain only the facts that are explicitly taken for granted, or should it also include other elements of the default background? According to Bratman [1991, p.30, my emphasis], in a particular deliberation, "if one has a *relevant* all-or-none, context-independent belief that p, and this belief is not bracketed, then one accepts p in that context. And similarly concerning a context-independent degree of confidence of 1." Observe that, without the emphasized notion of relevance, the adjusted cognitive background becomes rather large. It includes what is explicitly taken for granted and all the default background beliefs that can be added to it—consistently, as we shall see. The notion of relevant beliefs is aimed at limiting the size of the adjusted background. I return on the notion of relevance in Section 6.4.3. It is enough for now to see that the adjusted background might not decide on every single fact.

6.3.3 Requirements on the background of deliberation

There seem to be norms of rationality that apply to the constituents of the adjusted cognitive background, regardless of whether they are acceptances or beliefs. As we shall see, these general norms on the adjusted background mirror the norms

[26] The ideal agents that I have studied in the previous chapters do not, of course, have such constraints. I come back to this in Section 6.5, the conclusion of the present chapter.

on beliefs that I presented in Section 6.2.2. This can be seen as a consequence of the fact that beliefs and acceptances play essentially the same role in deliberation. Paraphrasing Velleman [2003, p.18], one can say that the adjusted cognitive background aims at picturing the world, of which there is just one, where the various courses of action considered shall be pursued. This can be seen as the starting point of an argument for the various norms that I am about to present. But I will not go into such a justification. My goal here is rather to present the norms on the adjusted background, from which follow most norms on acceptance, and to show how can one derive the norms on intentions from them.

Closure under logical operations. I first assume that the cognitive background of deliberations should be "logically" closed. I do not go into much detail about which logical rules of inference I mean here. For what follows I "only" suppose that the adjusted background should be closed under classical implication and conjunction. This means that if I accept in a given deliberation both that by doing A I also do B, and that I will do A, then I should also incorporate in my deliberation the fact that I will do B. Similarly, what an agent includes in the background should agglomerate. If ϕ and ψ feature in the adjusted background of a deliberation, then their conjunction should, too.

Obviously, agglomerativity of acceptances follows from this general closure requirement. If, in a particular deliberation, an agent takes ϕ for granted and takes also ψ for granted, he should take their conjunction for granted. But the agent is no more required to hold to this agglomerated acceptance in other contexts than he is with respect to the conjuncts. I take the same to apply when one of the conjuncts comes from a context-independent belief and the other from an acceptance. In this case the conclusion is also restricted to the given deliberation. Agents are required to carry agglomerated cognitive attitudes from one deliberation to another only when this operation can be done in the default background, as for example when two beliefs agglomerate[27].

To follow up on the remark at the end of Section 6.3.2, it is worth noticing that this closure requirement makes the adjusted cognitive background considerably larger. It does not, however, makes it complete in the sense of deciding on all facts.

Internal consistency. Beliefs were required to be internally consistent, and I assume that this is also the case in general for the elements of the adjusted cognitive background. Here I have in mind internal consistency in a strict sense: an agent should not include plain contradictions into the background of his deliberation. Internal consistency of acceptances is, of course, a special case of this general norm of consistency.

[27] This applies more generally to all "conclusions" that are reached via the closure requirement.

Another important consequence of internal consistency, together with the logical closure requirement, is a context-dependent form of *belief consistency* for acceptances. What is taken for granted should be consistent with the “believed” facts that are imported into the adjusted background. Otherwise, by the closure requirement, one would quickly obtain a new internally inconsistent acceptance. The strength of this new requirement, of course, depends on which beliefs are actually carried in the adjusted background. But even without being specific about which beliefs should be imported, this belief consistency requirement precludes an agent from bracketing flat-out beliefs that are explicitly invoked in practical reasoning. This seems to be in line with condition (3) of the definition of acceptances. An agent who is convinced that ϕ and is ready to use it in a particular context cannot rationally take for granted that ϕ is not the case in that same context.

Observe that one may require a stronger form of belief-consistency for acceptances, namely that they should be consistent with *all* non-bracketed flat-out and probabilistic beliefs of degree 0 or 1, regardless of whether they feature in the adjusted background or not. This requirement does not follow from internal consistency and closure of the adjusted background, but I do not need it to carry out the derivation of belief consistency of intentions, and so I leave it open whether one should include it in a theory of acceptances.

Explanatory coherence. Explanatory coherence of beliefs is mirrored in their standard of correctness (see Section 6.2.2). Agents were required to hold beliefs about evidence, but the adjusted background of deliberation also contains acceptance, the standard of correctness of which is different from that of beliefs. I shall thus use the following generalization of explanatory coherence. For a given element of the adjusted cognitive background, one should be able to find another element in the background that underpins its correctness.

For elements of the adjusted background that come from relevant beliefs, this new requirement boils down to the one presented above. Things are different for acceptances, though. Recall that the truth of their content or the fact that they make the agent better is necessary for their correctness. In terms of explanatory coherence of the adjusted background, it means that, for a given acceptance, one should be able to find other elements in the background that provide either evidence in favour of what is taken for granted or facts about the practical circumstances that motivate this very acceptance.

The cognitivist argument from explanatory to means-end coherence requires a specific assumption on beliefs about what one will do. Namely, these are to be explained by beliefs about what one intends. Now, the same will be required of elements of the adjusted background. I assume that if an agent accepts in a deliberation that he will do A he should also accept that he intends to A, or at least that he will later form the intention to A.

From the three general norms on the adjusted background thus follow agglomerativity, internal consistency, a form of belief consistency and explanatory coherence of acceptances. It should be stressed that these norms are all context-dependent. Acceptances, unlike beliefs, are not required to be consistent nor to agglomerate across contexts. I now investigate whether, with these characteristics to hand, one can explain internal consistency, belief consistency and means-end coherence of intentions without assuming that they involve beliefs.

6.4 Hybrid Pragmatism

Cognitivism is so called because it aims at showing that the norms of practical rationality associated with intentions come, in fact, from the norms of theoretical rationality associated with beliefs. Similarly, I have called Bratman's approach an "agency approach" because it tries to find an explanation for the norms of practical rationality in general structures of agency[28]. Here I want to see whether one can derive the norms on intentions from norms on acceptances that mirror the norms of theoretical rationality for beliefs. But this derivation is based on the idea that intentions involve acceptances, and I find that the most fitting overall justification for this requirement is a pragmatic one. So even though the approach I investigate here has an important cognitivist component, it is also firmly pragmatist. For that reason I call it hybrid pragmatism.

In this section I first present the assumption that drives hybrid cognitivism, namely that intentions "involve" acceptance. I then go on to say a few words on the notion of relevance for a particular deliberation, and come back to the absent-minded cyclist case. The core of the section is the last part, in which I finally look at how the norms on intentions can be derived from the norms on acceptance.

6.4.1 Intentions and acceptances

The key idea underlying the explanation of the requirements on intentions is the following:

(1) Having the intention to do A implies, in deliberations where this intention is relevant, accepting that one will do A.

This is indeed very close to the intermediate cognitivist assumption, with the crucial difference that beliefs are replaced by acceptances. One can find many statements that get close to (1) in the literature, notably in [Harman, 1976, p.438], [Bratman, 1999, p.32] and [Wallace, 2006, postscript to chap.5]. The changes from beliefs to acceptances, however, introduces a complication that the

[28] See the note on page 151.

notion of *relevance* for a deliberation tries to handle. The problem is that it does not seem realistic to assume that one can explicitly accept a very large number of facts when one deliberates[29]. The notion of relevance aims precisely at avoiding such an “overload” of acceptances. If, for example, I intend now to write my PhD thesis, there are many deliberations where I do not have to take that fact for granted. For most of my daily decisions this is just not relevant. But, as we shall see, relevance also plays a crucial role in the derivation of the requirements on intentions.

6.4.2 The absent-minded cyclist revisited

Before going further it is worth stressing that hybrid pragmatism is *not* bound to accept that the absent-minded cyclist described in Section 6.2.5 does not really intend to stop at the bookstore.

Recall the key features of this case. The agent intends to stop by the bookstore but he is uncertain whether he will in fact do so, because he knows that he tends to go on “automatic pilot” once on his bike. According to (1), he must come to accept that he will stop at the bookstore in further deliberations where this intention is relevant. But one can still deny that the agent would also assume that he will stop in theoretical reasonings. In short, one can still deny that the agent believes that he will stop at the bookstore.

6.4.3 Relevance for a deliberation

I have already mentioned twice the notion of relevance for a deliberation: in the Section 6.4.1, where it applied to intentions with respect to deliberations, and in Section 6.3.2, where it constrained the beliefs that have to be incorporated in the adjusted cognitive background.

These two uses of relevance regulate what should appear in the adjusted cognitive background. Indeed, in (1) it limits the intention-based acceptances that have to be incorporated in the adjusted background. Similarly, to say that beliefs, if not posited or bracketed, count as acceptances in deliberations where they are relevant is also to constrain what has to feature in the adjusted background. So, even though I have twice before mentioned the idea of relevance, relevance of beliefs once and relevance of intentions once, it served the same purpose in both cases.

I spell out the notion of relevance using the deliberately vague notion of “reasonably direct influence”. I state that a belief is relevant for a deliberation if the outcome of this deliberation depends in a reasonably direct way on the truth of the belief. In other words, if the content of the belief can make a reasonable difference in the deliberation, I state that this belief is relevant. The case is similar

[29] This, again, is not a problem for ideal agents. More on this in conclusion of this chapter.

for intentions. I state that an intention is relevant for a deliberation whenever the fact that one would execute the intention under consideration can in a reasonably direct way influence the outcome of the deliberation. The obvious case is when one takes into account the fact that the agent will accomplish what he intends influences the payoff structure of a deliberation as in, for example, situations of "overloading" [Pollack, 1991]. But it may also be that some options in a given deliberation enhance or diminish the feasibility of a pre-existing intention[30]. In that case the influence is less direct. The intention is relevant because the agent might want to take into account the consequences of his choice for the intentions that he has already settled on.

I do not think that it is necessary to get very precise about this notion of "reasonably direct influence". I put it forward only because it seems too strong to suppose, for example, that all beliefs whose content can have the slightest influence on the outcome of a deliberation are relevant. That would, I think, make too many beliefs relevant, and the same for intentions. On the other hand, there are cases where intentions or beliefs are obviously relevant for a deliberation; for example, when choosing one option in a deliberation would make it *impossible* to achieve some intention. In such a case it is clear that the deliberation influences the prospect of achieving the intention. The correct notion of relevance for a deliberations probably lies somewhere between these two extremes. But I do not think it is useful, nor easy, to draw a sharp distinction between what is relevant to a deliberation and what is not[31].

This is not to say that this notion of relevance is unimportant for what follows. Quite the contrary. On the one hand, it embodies a strong concern for the limited capacities of non-idealized planning agents by keeping down the number of facts that have to be taken into account during deliberation. But it nevertheless forces, so to speak, some acceptances to feature in deliberations. This provides hybrid pragmatism with a natural shield against the problem of false beliefs.

6.4.4 Rational intentions from rational acceptances

This section and the next are the keystones of this chapter. I look at how well hybrid pragmatism supports the derivation of internal consistency of intentions, strong belief consistency, means-end coherence and agglomerativity.

Internal consistency. This requirement is the easiest to derive. The argument is brief, and essentially parallels the one presented in Section 6.2.4. By (1), an internally inconsistent intention generates an internally inconsistent acceptance. Putting this contrapositive argument in the other direction, this means that internal consistency of acceptances implies internal consistency of intentions.

[30] See e.g. [Horty and Pollack, 2001].

[31] I should warn the reader that, in what follows, I often just use "influence", living implicit the "reasonably direct" proviso.

Strong belief consistency. The argument for belief consistency is more involved because one has to deal with the notion of relevance for deliberation. The difficulty lies in the fact that intentions have to be realizable in a world that corresponds to the *default* cognitive background. For the intermediate cognitivist this is not a problem because its main assumption goes directly from intentions to beliefs, which "live" in the default background, so to speak. But hybrid pragmatism stays at the level of the adjusted cognitive background of relevant deliberations. To carry the derivation through in this framework, one must show that, in case of strong belief inconsistency, the contradictory belief-intention pairs are somehow jointly relevant in certain deliberations. But once this is shown, the argument is more or less the same as in Section 6.2.4.

Suppose an agent has strongly belief-inconsistent intentions. He has the intention to do A but A cannot be realized in a world where his (default) background beliefs turn out to be true. Here I assume that this is the same as his believing that A cannot be done. Observe that this belief should be included in the adjusted cognitive background of a deliberation if this deliberation can in a reasonably direct way be influenced by the fact that A cannot be done. Take any such deliberation. Since doing A implies that A can be done, which is just the negation of something we assumed is relevant for this deliberation, we get that the fact that the agent will do A can also influence—through one of its consequences—the upshot of that deliberation. But this is just saying that the intention to do A is also relevant for that deliberation. This means, by (1), that the agent should also include in the cognitive background of that deliberation the fact that he will do A. But this new acceptance is belief inconsistent[32].

Agglomerativity. Let me once again begin with agglomerativity against potential irrationality. As for cognitivism, hybrid pragmatism derives it automatically given that it can explain internal and strong belief consistency. Again, there are two cases to consider. Since they are essentially similar, I only sketch the argument for agglomerativity against potential violation of belief consistency. Suppose that an agent has the intention do to A and the intention do to B which, were they conjoined in a single intention, would not be achievable, given his beliefs. We know from the previous section that this hypothetically conjoined intention

[32] The reader should also bear in mind that internal consistency and agglomerativity of the adjusted cognitive background are at work here. They were used to derive belief consistency of acceptances. It should also be noted that the argument could also have proceeded as follows. Take any deliberation on means to achieve the intention to do A. It is clear that the fact that A cannot be done is relevant to such deliberation, and so that it should feature in the adjusted cognitive background. Given (1), one thus directly obtains a violation of beliefs consistency of acceptances. This argument is not as general as it should be, if one grants that some intentions never require any further deliberations on means. I honestly doubt that there are such intentions. But in the absence of an explicit argument for that claim, I think that hybrid pragmatism should retain the more abstract derivation presented above.

would generate a violation of the belief consistency of acceptances, which is just what was needed.

Agglomerativity as closure, however, is more difficult to derive. Acceptances are indeed agglomerative *within* relevant backgrounds of deliberations. But here this restriction to the adjusted background is a blight rather than an asset. For suppose that an agent intends to do A and intends to do B. Why should he intend to do A and B? To use agglomerativity of acceptances, we would have to make sure that these two intentions are relevant to at least one common deliberation. But since we are considering arbitrary intentions, I do not see why this should be so.

We have seen in Section 6.2.4 that strong cognitivism can, however, easily justify this requirement. Maybe a stronger form of hybrid pragmatism, characterized as follows, could work.

(2) An agent intends to do A if and only if he accepts, in relevant deliberations, that he will do A.

At first sight, such strong hybrid pragmatism is tempting, because acceptances are functionally very close to intentions. But just like its cognitivist analogue, (2) comes with independent problems. Namely, it makes it difficult to distinguish unintended side effects from intended consequences of actions, as argued in [Bratman, 2006b, p.18-20]. For that reason I think that one should resist the temptation to accept (2) and rather stick to (1). This, of course, means that hybrid pragmatism still falls short of an explanation of agglomerativity as closure of intentions. Given that this principle can itself be questioned, as we saw in Section 6.1, this might not be a very serious problem.

It should be noted, however, that if one adopts a principle of agglomerativity as closure restricted to plans, relevance in deliberation might come back into force and provide what we need to push agglomerativity as closure. For one would no longer be dealing with arbitrary intentions, but rather with intentions that are part of a single plan. This could help to find the common relevant deliberations that are absent in the general case.

I shall not go in that direction here, though. It would require me to specify the adequate level of agglomeration, and I have for now no precise idea of how one could do this. It is telling, however, that the only argument I sketched for agglomerativity as closure with a restricted scope was, in essence, pragmatic (see Section 6.1). I mentioned that, at the level of plans, to systematically agglomerate intentions might help personal coordination in extensive decision problems. If this argumentative strategy turns out to be fruitful, it would in itself lobby in favour of the "hybrid" character of the derivation using acceptances.

Means-end coherence. The main ingredient in the derivation of means-end coherence of intentions is explanatory coherence of acceptances. But, just like cognitivism, using means-end coherence makes hybrid pragmatism vulnerable to the problem of false beliefs. I shall return to this at the end of the section. There is, in the meantime, another complication that the hybrid pragmatist derivation has to overcome.

This complication comes once again from the fact that (1) only requires the agent to take things for granted in the adjusted background of relevant deliberations, while the belief that "triggers" means-end incoherence is primarily a denizen of the default background. The main step of the derivation is thus to show that the means-end incoherent belief-intention pair can be relevant to at least some common deliberations. Just as for belief consistency, once this is shown the argument is fairly similar to the one in Section 6.2.4.

Suppose that an agent intends to do E and believes that he will achieve E only if he does (M_1 or M_2 or... or M_n). Take a deliberation the upshot of which can affect the prospect of achieving E. As noted in Section 6.3.3, this is a case where the fact that the agent will do E can influence the upshot of that deliberation. The agent might want to take into account the effect of his decision on his background plans. This means, first, that this intention is relevant for that context, and so by (1) that he should include in the adjusted background the fact that he will do E. Observe, however, that the upshot of the deliberation can affect the prospect of doing E only by affecting the prospect of doing M_1 or M_2 or ... or M_n, at least according to what the agent believes. That is, the agent might want to take into account the effect of his decision on the feasibility of each of these means. This makes the fact that doing E implies doing one of these M is also an important input into that deliberation, which is just to say that it is also relevant here. But then the adjusted background features both the facts that the agent will do E and that doing E implies doing one of the Ms. By the closure principle for the adjusted background, the agent must come to accept that he will do M_1 or that he will do M_2 or... or that he will do M_n. Finally, by applying explanatory coherence one can conclude that the agent should also accept in that background that he intends one of these means or, at least, that he will come to intend one of them later.

As Wallace puts it in the context of the cognitivist derivation, this is as far as explanatory coherence of acceptances can get the derivation. But this is not quite as far as one needs, unless one can show that the agent should not take for granted that he has the required means-intentions without, in fact, having these intentions. In other words, mistaken acceptances about one's own intentions seem to threaten the hybrid pragmatist derivation just as it did the cognitivist one.

As we saw in Section 6.2.4, Wallace has proposed a solution that seemed to cope with the problem. This solution, however, involves a step outside the pure cognitivist perspective because it crucially invokes a principle of practical rationality. But hybrid pragmatism is built on the notion of acceptances in deliberations, which are sensitive to practical concerns. For that reason, I think it can naturally claim Wallace's reply to the problem of false belief, in a way that might be even more suited to his "hybrid" nature.

Recall that the essence of Wallace's reply was that, *in relevant deliberations*, it is independently irrational, *for general pragmatic reasons*, to hold false beliefs about one's own intentions. *Mutatis mutandis*, one can say that it is independently irrational, for general pragmatic reasons, to mistakenly take for granted one's own intentions in deliberations where the latter would be relevant.

I have already explained in Section 6.2.4 why Wallace's emphasis on relevant contexts blocks counterexamples like the one proposed by Bratman. The explanation is readily applicable to acceptances, and is arguably even more convincing in these terms[33].

But to push through the derivation of means-end coherence one needs not only to show that this particular counterexample could be handled; one needs to show that, in general, mistaken acceptances about one's own intentions are irrational. This was the object of Wallace's recourse to a general principle of practical rationality. Hybrid pragmatism is, once again, especially suited to the incorporation of such a principle. Acceptances are after all responsive to practical concerns.

The reader should observe, however, that Wallace's general type of practical concerns is not the same as the one that features in the characterization of acceptances (Section 6.3.1). The latter is specific to deliberation contexts, while the former are "to be included among the traits and capacities that make us, *in general*, effective in the pursuit of our goals". In fact, there is a principle of this sort that can be used to provide a "wide scope" justification of (1), which I shall look at in the next section[34].

Before I do that, though, let me summarize where things stand now. Hybrid pragmatism can easily explain internal consistency of intentions and, after taking care of complications regarding relevance in deliberation, strong belief consistency as well as agglomerativity against potential irrationality. Just like intermediate cognitivism, however, it fails to explain agglomerativity as closure. It can also explain means-end coherence, provided that we can secure the irrationality of

[33] This reply also takes care of another potential counterexample. Acceptances are, as we saw, both responsive to evidence and practical concerns. Now, suppose that an agent has, in the default background, a mistaken belief about his own intention, which is supported by some misleading evidence. Would it not then be rational of him to incorporate this fact in relevant deliberation? The answer, I think, should be "no", because such deliberations are again precisely deliberations where the fact that the agent has the intention is relevant.

[34] To the extent that the last step of the argument succeeds, it shows how suited to hybrid pragmatism is Wallace's attempt to save the cognitivist derivation of means-end coherence. Except for the switch from acceptances to beliefs, he spelled out all the key steps of the argument. He saw the importance of looking at deliberation where one's own intentions are relevant, and connected it to pragmatic consideration. In fact, his suggestion is much more at odds with the orthodox cognitivist line than with hybrid pragmatism. At the end of the day, one may wonder why he did not directly speak in terms of acceptances.

mistaken acceptances about one's own intentions. This requires a more general pragmatic argument, which I now explore.

6.4.5 Acceptances and intentions in deliberation

The hybrid pragmatist derivation of the norms on intentions rests on two assumptions that still have to be justified: the implication (1) from intention to acceptances and the connection between intention and acceptances about intentions. In this section I first sketch how one can provide a two-step argument for (1). At the core is a constitutive claim in favour of (1), which is cloaked by a "wide scope" pragmatic justification. The ideas driving this second step are quite unifying, allowing one also to secure the derivation of means-end coherence. I finally look back at the constitutive claim, in the light of a more general worry about the hybrid pragmatist enterprise.

A key ingredient in the theory of practical reasoning for planning agents that I have developed in this thesis is that intentions should not only be seen as outputs of deliberation, but also as important inputs. This influence of previously adopted intentions and plans on deliberation is twofold. They trigger deliberation about the means by which they will be achieved and they rule out of consideration options with which they are incompatible.

The planning theory, however, says little about *how* they do so, and it seems that acceptances help to understand this connection. More precisely, the implication (1) from intentions to acceptances, together with the various norms on the latter, provide the required connection between intentions and deliberation.

Let me first look at the filtration of options that comes from intentions. How is that supposed to happen? Bratman [1991, p.32] proposed that "what is required of an option in deliberation is that one's so acting be consistent with one's intentions and beliefs, and that one accept in that context that one will so act if one so decides". One can say, more generally, that options in deliberations have to be consistent with the adjusted cognitive background. But then, if we assume (1), the options to be considered in a deliberations will have to be consistent with the agent doing what he intends. Conversely, (1) gives us a way to see why intentions rule out inadmissible options. They do so via acceptances.

The situation is analogous with respect to the standard of relevance with respect to means that intentions impose. Recall the reasoning that I pictured in Section 6.4.4. I started with an intention to do E and the belief that to do E one must do (M_1 or M_2 or... or M_n). From these I concluded, after a few intermediate steps which crucially involved (1), that one should accept that he will do M_1 or that he will do M_2 or... or that he will do M_n. But, in the absence of a mistaken acceptance that the agent already has the corresponding intention, recognizing that this intention is absent may well trigger means-reasoning. In other words, becoming aware of this threat of means-end incoherence may trigger deliberation about means. Note that this is not to say that the agent should form the intentions required to resolve means-end incoherence. For now the claim is only constitutive. With (1), intentions about ends can find their way into the adjusted background and set off deliberation about means.

So (1) can be seen as a way to understand how intentions influence practical reasoning. As I just mentioned, this is a constitutive claim. The next step is to ask why the intentions of planning agents *should* play this role. In other words, is there a normative reason to think that planning agents should take their previously adopted intentions into account while deliberating? Here I think that the most promising answer to this question is the general pragmatist approach proposed by Bratman [2006b] and Wallace [2006][35]. Intentions generally "aim at coordinated control of action that *achieve what is intended*" [Bratman, 2006b, p.33, emphasis in original] or, in other words, intentions are "among the traits and capacities that makes us, in general, effective in the pursuit of our goals"[Wallace, 2006, p.118]. Forming intentions is, in short, a generally useful tool to achieve our goals. Many features of intentions make them useful. They are "conduct controlling", relatively resistant to reconsideration and, what is important for hybrid pragmatism, they influence further deliberations.

This normative claim indeed very well suits the hybrid nature of (1). If the reason why intentions should play a role in practical reasoning is a general, pragmatic one, it is natural to think that they will do so in a way that is also responsive to practical concerns, i.e. via acceptances. But one should bear in mind that it is a "wide scope", normative claim. In the words of Broome [1999], the "should" is not detachable. If an agent intends to do A it does not follow that he automatically has good, practical reasons to take it for granted that he will do A, especially if he did not have good reasons to form that intention to start with. Rather, the argument aims at showing that there are general pragmatic reasons for planning agents to comply with (1)[36].

This general pragmatic claim reaches further than the justification of (1). It also allows us to overcome the problem of mistaken acceptances. In exactly the same fashion as Wallace, one can say that to mistakenly accept that one intends something in relevant deliberation is irrational because it threatens the general efficiency of planning agents to reach their goal. Again, bluntly, mistaken acceptances can mess up the agent in the long run[37]. With this in hand, hybrid pragmatism has a way to explain the means-end coherence requirement for intentions, in a way that fits the overall justification of the others requirements. As such, it seems to provide a more unified argumentation than Wallace's sup-

[35] Observe that one is not bound to take this route. Since (1) is taken here as a constitutive claim, one could as well try to cloak it with a more cognitively-oriented normative argument. I choose the pragmatist one because it seems to me the more plausible.

[36] In quasi-formal terms, the argument aims at showing that "pragmatic-should (Intention A → Acceptance that A)" and not that "Intention A → pragmatic-should (Acceptance that A)".

[37] In fact, one could argue that they can mess up even more than mistaken beliefs, because of the crucial context-dependency of acceptances.

plemented cognitivism[38]. In this case one could wonder why one suddenly needs to appeal to practical concerns to derive means-end coherence, while this is not the case for belief and internal consistency. But if one uses acceptances, practical concerns are there all along.

This raises a more general question about the motivations behind the hybrid pragmatist view. One attractive feature of cognitivism, to the extent that it succeeds, is its clear stance: the practical norms on intentions can be explained only in terms of theoretical norms on beliefs. The same can be said about Bratman's agency approach: the practical norms on intentions can be explained only in practical terms[39]. In comparison, hybrid pragmatism sits in a rather grey area, invoking both pragmatic and theoretical-like requirements.

I think one can better appreciate the relevance of such a mixed approach by looking at the constitutive claim mentioned at the beginning of this section. Intentions are playing a role in practical reasoning via their influence on the *cognitive* background of deliberation, which is under some pressure to represent accurately the world. From that point of view, it seems that the norms on intentions fall in two categories. On the one hand, the derivations of agglomerativity against potential irrationality, of internal consistency and of strong belief consistency rest on the requirements of closure under logical operation and of internal consistency for the adjusted background of deliberation. These norms obviously mirror norms of theoretical rationality. This should not come as a surprise. To paraphrase again Velleman [2003], the adjusted cognitive background seems to aim at picturing the world, of which there is just one, where the various courses of action considered shall be pursued. It plays in practical reasoning a role similar to the role of the default background in theoretical reasoning. Inasmuch as the argument in Section 6.4.4 is correct, the norms of agglomerativity against potential irrationality, of internal consistency and of strong belief consistency ultimately derive from this role of the adjusted background. This gives a strong "cognitive" character to these norms, by which they somehow differ from means-end coherence. The derivation of this norm on intention indeed crucially rests on practical concerns, especially to avoid the problem of false beliefs or of mistaken acceptances. As such, means-end coherence appears to be, unlike the three other norms just mentioned, a genuinely practical norm on intentions.

This general distinction between the practically- and the cognitively-oriented norms on intentions help to understand why "pure" cognitivism as so much difficulty with means-end coherence, and why it seems forced in the end to fall back on pragmatic concerns. To explain this practically-oriented norm one needs at least some pragmatist component. But, on the other hand, a pure agency approach seems to leave behind the fact that some norms on intentions have a

[38] The epithet comes from Bratman [2006b, p.8].

[39] One should take care about where the "only" is located. Bratman does not seems to claim that the practical norms can only be explained in practical terms.

important cognitive character. The theoretical-like constraints that they embody are perhaps better explained in their relations with the cognitive background of deliberation, which is precisely what hybrid cognitivism does by using acceptances.

6.5 Conclusion

Hybrid pragmatism is thus a plausible philosophical approach to the norms of consistency and coherence which apply to intentions. It not only helps us to understand where these norms come from, but also allows us to explain how intentions influence deliberation, via the notion of acceptances. That this approach definitely stands in between cognitivism and Bratman's agency approach can also be seen as an asset. It provides an account that does justice to the influence of intentions on both the pragmatic and the cognitive component of practical reasoning. In other words, it provides a unifying background to the role of intentions in rational decision making, something that was arguably missing in the last chapters.

To conclude this chapter I would like to address one final question, bearing to resource-boundedness[40]. As I explained in Chapter 2, I have carried the formal investigations under strong assumptions about the capacities of agents: all issues pertaining to resource-boundedness were ignored. Resource-boundedness played, however, an important role in the discussion of hybrid pragmatism. A crucial asset of acceptances is that they simplify deliberation, by for instance provisionally settling issues that the agents is uncertain about. Resource-boundedness also came back through the idea of relevance for deliberation, in Section 6.4.3, and during the discussion of the problem of false beliefs (Sections 6.2.4 and 6.4.4). Does it mean that hybrid pragmatism is a suitable theory of how intentions influence practical reasoning for limited agents, but not for the type of agents that I studied in Chapters 2 to 5?

I think not, for the following reasons. First, in the discussion around the idea of relevance for deliberation and around the problem of false beliefs, resource-boundedness rather raised difficulties for the hybrid pragmatist justification. It has, for instance, forced to look carefully at what agents should accept in deliberation, precisely for the reason that resource-bounded agents can only take for granted a limited amount of facts. Since this problem does not arise for non-resource-bounded agents, the hybrid pragmatist derivation of the norms on intention seems even more straightforward under the idealizations I made in Chapters 2 to 5. Second, the idea that acceptances are of greater importance for agents with limited capacities does not undermine their explanatory power in the general case. They still provide a simple way to understand how intentions influence practical

[40] I am grateful to Michael Bratman (p.c.) for pointing me out this potential weakness in the connection between the present chapter and the foregoings.

reasoning—the "constitutive claim" of Section 6.4.5—an issue that has been very little addressed in the philosophical literature. In other words, acceptances in deliberation and hybrid pragmatism complete the picture of how future-directed intentions constraint practical reasoning, whether for ideal or resource-bounded agents. As such, the philosophical theory developed in this chapter seems indeed to encompass both the case of agents with limited capacities and the ideal agents which have been studied in Chapters 2 to 5.

Of course, hybrid pragmatism also opens up questions about intention-based practical reasoning from a formal perspective. How is one to include acceptances in the formal theory of deliberations that I proposed? How will they be distinguished from other kinds of "informational" states, like knowledge and belief? How, in these models, can one account for the fact that acceptances are also responsive to pragmatic considerations? These are crucial questions that a formal enquiry would surely help answer. Furthermore, one can hope that bringing acceptances into the formal picture will also unveil new issues about them, just as in the previous chapters game and decision-theoretic models unveiled new issues about intentions. All in all, hybrid pragmatism is very well suited as the final theme for this thesis. It rounds up issues that were carried all the way, and opens up exiting new ones.

Chapter 7

Conclusion

In this thesis I have proposed a theory of practical reasoning which drew on three contemporary paradigms: instrumental rationality from decision and game theory, epistemic reasoning from philosophical logic and computer science, and planning agency from philosophy of action. This provides a unified theory of rational planning agency, which is a theory of how agents deliberate when they take into account the demands of instrumental rationality, their background of future-directed intentions and the information they have about the rationality, intentions and information of others.

7.1 Review of the chapters

I have shown in Chapter 2 that such a broad perspective can account for personal coordination in extensive decision problems, because rational planning agents are able to break ties between equally desirable options. In Chapter 3 I brought this tie-breaking effect to the level of interactive situations, and I have shown that it provides a natural anchor for interpersonal coordination in Hi-Lo games. With the help of epistemic models for these games, I have been able to study explicitly how mutual knowledge of intentions also foster coordination in games. By the same token I was able relate the intention-based explanation of coordination to other accounts in the game theoretical literature, and to circumscribe better the differences between coordination and fully cooperative shared agency.

In Chapter 4 I have studied how rational planning agents transform their decision problem on the basis of what they intend, a phenomenon called the reasoning-centered commitment in philosophy of action. I have shown that this aspect of intention-based practical reasoning is especially sensitive to interactions. When many planning agents simultaneously transform the decision problem they face, it becomes crucial that they take each others' intentions into account.

Chapter 5 merged the considerations of the previous chapters into a unified picture of rational planning agency. Using dynamic epistemic logic, I have been able to relate the informational components of intentions-based deliberation with the active process of decision problem transformation. I have also shown that this framework really does capture phenomena at the intersection of planning agency and instrumental rationality, such as the relation between the game-theoretic solution concept of elimination of dominated strategies and the filtering of intention-inconsistent options. Finally, I have provided this framework with axiomatic proof systems, which give an explicit representation of practical reasoning in games with intentions.

In Chapter 6 I explored the philosophical basis of this theory of intention-based deliberation, looking at where the various norms of consistency and coherence of intentions come from. This led to hybrid pragmatism, an attempt to explain these norms in terms of similar norms which apply on acceptances in deliberation. I have argued that hybrid pragmatism is a plausible alternative to the main contemporary proposals in philosophy of action, because it does justice to both the cognitive and the pragmatic side of the norms of consistency and coherence. Furthermore, it provides a natural explanation of how future-directed intentions influence practical reasoning, and as such helps us to see better how the various pieces encountered in this thesis fit together.

7.2 Open questions

Of course, many questions were left unanswered along the way. Instead of reviewing them, I will rather present three broad research themes that this thesis opens up. The first relates more specifically to logic, the second to the theory of intentions and the third to game theory. I think, however, that from a general point of view each of them is relevant to all these fields.

I have many times mentioned that the present framework is in great need of a more elaborate theory of intention revision. This poses challenging problems from a logical point of view. Dynamic epistemic logics for *belief* revision have been extensively studied[1]. These systems have very interesting logical properties, in terms of axiomatizations and expressive power, and it is surely worth looking at how they would transfer to logic for intention revision. What it more, developing the logic of intention revision is surely a good way to establish the connection between the approach adopted in this thesis and BDI architectures. As I mentioned at the end of Chapter 5, BDI models are the main contemporary logical approach to intention-based practical reasoning. The road I have taken here did not quite allow for a systematic comparison, but at this point the issue is, to say the least, pressing.

From the philosophical point of view, hybrid pragmatism and the theory of acceptances in deliberation that I used in Chapter 6 are “new” issues that deserve

[1] Girard [2007], van Benthem [2007] and Baltag and Smets [2006] are especially interesting from the current perspective.

much more detailed scrutiny. But the investigation that I carried out in Chapter 4 has also unveiled, I think, an important gap in the "core" theory of intentions. The intricacies of intention-based transformation of decision problems in interaction situations have been widely overlooked in the philosophical literature[2]. In comparison, the notion of individual intentions with a "we content" that I used in Chapter 3 has attracted much attention, and raised important questions about conditions under which agents are *justified* in forming them. Similar questions obviously also apply to intention-based transformations of decision problems. Are agents always justified in cleaning or pruning their option set, even when they are uncertain about the intentions of others? If not, can one phrase the appropriate conditions of justification in terms of mutual knowledge of intentions, as it is done for intentions with a "we" content? These issues are of the greatest importance for the theory of intentions, because they concern the very "regularities which connect intentions with each others, with associated processes and activities, and with characteristic 'inputs' and 'outputs'." [Bratman, 1987, p.9] In short, they concern what intentions *are*.

The place of rational deliberation with intentions in interactive situations also poses very interesting problems from the point of view of game theory. As mentioned in Chapter 3, the intention-based account of coordination in Hi-Lo games occupies a point somewhere in the middle ground between purely competitive and genuinely cooperative scenarios. It makes crucial use of intentions of the form "I intend that we..." which, even though they are essentially individual states, have a clear social character. As such, intention-based practical reasoning is a bridge between competitive and cooperative game theory. It thus offers a plausible alternative to the Nash program [Serrano, 2005], a well-known attempt to translate cooperative into non-cooperative frameworks.

D. Davidson's seminal contribution [1980] to contemporary philosophy of action was profoundly influenced by his familiarity with models of instrumental rationality from theoretical economics, and especially with decision theory [Malpas, 2005]. Since then, however, these disciplines have mostly evolved in parallel, and only recently can we see a renewal of interest in genuinely interdisciplinary work on rational decision making, witness e.g. the work of Parikh [2002], van Benthem [2006] and Bacharach [2006]. I have written this thesis with the conviction that such interdisciplinary approaches are fruitful, and my hope is that I have conveyed this conviction to the reader.

[2] Harp [2008] is a notable exception.

Bibliography

M. Aiello, J. van Benthem, and G. Bezhanishvili. Reasoning about space: The modal way. *Journal of Logic and Computation*, 13:889–920, 2003.

F. Alonzo. *Shared Intention, Reliance, and Interpersonal Obligations.* PhD thesis, Stanford University, forthcoming.

F.J. Anscombe and R.J. Aumann. A definition of subjective probability. *Annals Math. Stat.*, 34:199–205, 1963.

G.E.M. Anscombe. *Intention.* Harvard University Press, Harvard University Press, 1957.

K.R. Apt. The many faces of rationalizability. *The B.E. Journal of Theoretical Economics*, 7(1), 2007. Article 18.

Aristotle. *Nichomachean Ethics.* Oxford University Press, 1962. Translation M. Ostwald.

K.J. Arrow. *Social Choice and Individual Values.* Yale University Press, 1970.

R.J. Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, 55(1-18), 1987.

R.J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:121–133, 1994.

R.J. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28:263–300, 1999.

R.J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976. URL `http://links.jstor.org/sici?sici=0090-5364%28197611%294%3A6%3C1236%3AATD%3E2.0.CO%3B2-D`.

R.J. Aumann and A. Brandenburger. Epistemic conditions for nash equilibrium. *Econometrica*, pages 1161–1180, 1995.

R.J. Aumann and S. Sorin. Cooperation and bounded recall. *Games and Economic Behavior*, 1(1):5–39, 1989.

J.L. Austin. How to talk: Some simple ways. *Proceedings of the Aristotelian Society*, 53:227–246, 1953.

M. Bacharach. *Beyond Individual Choices: Teams and Frames in Game Theory.* Princeton University Press, Princeton, 2006. Edited by N. Gold and R. Sugden.

A.C. Baier. Act and intent. *The Journal of Philosophy*, 67(19):648–658, 1970.

A. Baltag and S. Smets. Conditional doxastic models: A qualitative approach to dynamic belief revision. In G. Mints and R. de Queiroz, editors, *Proceedings of WOLLIC 2006, Electronic Notes in Theoretical Computer Science*, volume 165, 2006.

A. Baltag and S. Smets. Knowledge, safe belief and dynamic rationality: logics for higher-order belief revision. Oxford and Vrije Universiteit Brussel, Unpublished manuscript.

A. Baltag, L.S. Moss, and S. Solecki. The logic of public announcements, common knowledge and private suspicions. In *TARK 98*, 1998.

K. Binmore. *Fun and Games: A Text on Game Theory.* Houghton Mifflin College Div, 2005. 2nd Edition.

P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic.* Cambridge University Press, Cambirdge, 2001.

P. Blackburn, J. van Benthem, and F. Wolter, editors. *Handbook of Modal Logic.* Elsevier, November 2006.

G. Bonanno. Two lectures on the epistemic foundations of game theory. URL `http://www.econ.ucdavis.edu/faculty/bonanno/wpapers.htm`. Delivered at the Royal Netherlands Academy of Arts and Sciences (KNAW), February 8, 2007.

C. Boutilier. Toward a logic of qualitative decision theory. *Proceedings of the 4th Intl. Conf. on Principle of Knowledge Representation and Reasoning (KR-94)*, 1994. URL `http://www.cs.toronto.edu/~cebly/Papers/kr94.ps`.

A. Brandenburger. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35:465–492, 2007.

A. Brandenburger and E. Denkel. Rationalizability and correlated equilibria. *Econometrica*, 55:1391–1402, 1987.

M. Bratman. *Structures of Agency: Essays.* Oxford University Press, 2006a.

M. Bratman. Intention, belief, practical, theoretical. Unpublished Manuscript, Stanford University, January 2006b.

M. Bratman. *Intention, Plans and Practical Reason.* Harvard University Press, London, 1987.

M. Bratman. *Faces of Intention; Selected Essays on Intention and Agency.* Cambridge University Press, 1999.

M. Bratman. Practical reasoning and acceptance in a context. *Nous*, 1991. Reprinted in [Bratman, 1999, p.15-34] The page numbering in the various quotes refers to this second print.

M. Bratman, D.J. Israel, and M.E. Pollack. Plans and resource-bounded practical reasoning. In J. Pollock and R. Cummins, editors, *Philosophy and AI: Essays at the Interface*, pages 7–22. MIT Press, 1991.

J. Broome. Practical reasoning. In J. Bermudez and A. Millar, editors, *Reasons and Nature: Essays in the Theory of Rationality.* Oxford University Press, 2002.

J. Broome. Normative requirements. *Ratio (new series)*, XII(4), December 1999.

J. Brunero. Two approaches to instrumental rationality and belief consistency. *Journal of Ethics & Social Philosophy*, forthcoming.

D.K. Chan. A not-so-simple view of intentional action. *Pacific Philosophical Quarterly*, 80:1– 16, 1999.

B. Chellas. *Modal Logic: An Introduction.* Cambridge University Press, Cambridge, 1980.

L.J. Cohen. Belief and acceptance. *Mind*, 98(391):367–389, July 1989.

P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artif. Intell.*, 42(2-3):213–261, 1990. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/0004-3702(90)90055-5.

A. M. Colman and M. Bacharach. Payoff dominance and the stackelberg heuristic. *Theory and Decision*, V43(1):1–19, 1997. URL `http://dx.doi.org/10.1023/A:1004911723951`.

M. Cozic. *Fondements cognitifs du choix en rationalité limitée.* PhD thesis, Paris IV - Sorbonne, December 2005.

D. Davidson. *Essays on Actions and Events.* Clarendon Press, Oxford, 1980.

B. de Bruin. *Explaining Games : On the Logic of Game Theoretic Explanation.* Illc dissertation series ds-2004-03, Universiteit van Amsterdam, 2004.

D. de Jongh and F. Liu. Proceedings of the workshop on rationality and knowledge. In S. Artemov and R. Parikh, editors, *Optimality, Belief and Preference.* ESSLLI '06, 2006.

P. Engel. Believing, holding true, and accepting. *Philosophical Explorations*, 1 (2):140–151, 1998.

R. Fagin, J.Y. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge.* MIT Press, 1995.

K. Faley. Knowledge in intention. *Philosophical Studies*, 99:21– 44, 2000.

G. Gargov and V. Goranko. Modal logic with names. *Journal of Philosophical Logic*, 22:607–636, 1993.

M. Georgeff, B. Pell, M.E. Pollack, M. Tambe, and M. Wooldridge. The belief-desire-intention model of agency. In J. Muller, M. Singh, and A. Rao, editors, *Intelligent Agents V.* Springer, 1998.

J. Gerbrandy. *Bisimulations on Planet Kripke.* PhD thesis, ILLC, Amsterdam, 1999.

J. Gerbrandy, J. van Benthem, and E. Pacuit. Merging frameworks for interaction : DEL and ETL. In *Proceedings of TARK 2007*, 2007.

G. Gigerenzer and R. Selten. *Bounded Rationality: The Adaptive Toolbox.* MIT Press, 2002.

P. Girard. *Modal logics for belief and preference change.* PhD thesis, Stanford University, 2008.

P. Girard. From onion to broccoli: Generalizing lewis's counterfactual logic. *Journal of Applied Non-classical Logics*, 17(2), 2007.

H.P. Grice. Intention and uncertainty. In *Proceedings of the British Academy*, volume 57, London, 1971. Oxford University Press.

J.Y. Halpern. Defining relative likelihood in partially-ordered preferential structure. *Journal of Artificial Intelligence Research*, 7:1–24, 1997.

H.H. Hansen. Monotonic modal logics. Master's thesis, ILLC, Universiteit van Amsterdam, October 2003.

S.O. Hansson. Preference logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic (Second Edition)*, volume 4, chapter 4, pages 319–393. Kluwer, 2001.

G. Harman. *Change in View.* MIT Press, 1986.

G. Harman. Practical reasoning. *Review of Metaphysics*, 29(3):431–463, 1976.

R. Harp. *Collective reasoning, Collective action.* PhD thesis, Stanford University, 2008.

J.C. Harsanyi. Games with incomplete informations played by 'bayesian' players. *Management Science*, 14:159–182, 320–334, 486–502, 1967-68.

J. Hintikka. *Knowledge and Belief: An Introduction to the Logic of Two Notions.* ornell University Press, Ithaca, N.Y., 1962.

R. Holton. Deciding to trust, coming to believe. *Australian Journal of Philosophy*, 72(1), March 1994.

J.F. Horty and M.E. Pollack. Evaluating new options in the context of existing plans. *Artificial Intelligence*, Volume 127(2):199–220, April 2001.

R. Jeffrey. *The Logic of Decision.* McGraw-Hill, New-York, 1965.

P. Jehiel and D. Samet. Valuation equilibria. *Game Theory and Information 0310003, Economics Working Paper Archive at WUSTL*, 2003.

R. Johnson. The stanford encyclopedia of philosophy, February 2004. URL `http://plato.stanford.edu/`. E. Zalta (ed.).

J.M. Joyce. Bayesianism. In A.R. Mele and P. Rawling, editors, *The Oxford Handbook of Rationality.* Oxford University Press, 2004.

I. Kant. *Fundamental Principles of the Metaphysic of Morals.* http://eserver.org/philosophy/kant/metaphys-of-morals.txt, 1785. Translation T.K. Abbott, 2007.

G.S. Kavka. The toxin puzzle. *Analysis*, 43(1), 1983. ISSN 1467-8284.

N. Kolodny. How does coherene matters? Downloaded from the author's website on April 23th 2007, April 2007.

D.M. Kreps and R. Wilson. Sequential equilibria. *Econometrica*, 50:863–894, 1982.

P. La Mura and Y. Shoham. Conditional, hierarchical, multi-agent preferences. In *Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge*, pages 215 – 224, 1998.

C. List. Judgment aggregation, a bibliography on the discursive dilemma, doctrinal paradox and decisions on multiple propositions, 2007. http://personal.lse.ac.uk/LIST/doctrinalparadox.htm.

F. Liu. *Changing for the Better: Preference Dynamics and Agent Diversity.* PhD thesis, Universiteit van Amsterdam, February 2008.

D. R. Luce and H. Raiffa. *Games and Decisions; Introduction and Critical Survey.* Dover Publications, Inc., 1957.

J. Malpas. The stanford encyclopedia of philosophy, May 2005. URL `http://plato.stanford.edu/`. E. Zalta (ed.).

E.F. McClennen. *Rationality and Dynamic Choice : Foundational Explorations.* Cambridge University Press, 1990.

A.R. Mele, editor. *The Philosophy of Action.* Oxford University Press, 1997.

D. Mihalache. Safe belief, rationality and backwards induction in games. Master's thesis, Oxford University Computing Laboratory, September 2007.

R.B. Myerson. *Game Theory: Analysis of Conflict.* Harvard University Press, 1997 edition, 1991.

J. Nash. Equilibrium points in n-persons games. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 36, pages 48–49, 1950.

M.J. Osborne. *An Introduction to Game Theory.* Oxford University Press, New-York, 2004.

M.J. Osborne and A. Rubinstein. *A Course in Game Theory.* MIT Press, 1994.

B. O'Shaughnessy. Trying (as the mental "pineal gland"). *The Journal of Philosophy*, 70(13, On Trying and Intending):365–386, Jul. 19 1973.

E. Pacuit. *Neighborhood Semantic for Modal Logic. An introduction*, July 2007. Course notes for ESSLLI 2007.

R. Parikh. Social software. *Synthese*, 132(3), September 2002.

J.A. Plaza. Logics of public communications. In M.L. Emrich, M.S. Pfeifer, M. Hadzikadic, and Z.W. Ras, editors, *Proceedings of the Fourth International Symposium on Methodologies for Intelligent Systems: Poster Session Program*, pages 201–216. Oak Ridge National Laboratory, 1989.

M.E. Pollack. Overloading intentions for efficient practical reasoning. *Nous*, 25 (4):513–536, 1991.

M.E. Pollack. The uses of plan. *Artificial Intelligence*, 1992.

F.P. Ramsey. Truth and probability. In R.B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*. Routledge, 1926.

H. Rott. *Change, Choice and Inference: A Study of Belief Revision and Nonmonotonic Reasoning.* Oxford Logic Guides. ford University Press, Oxford, 2001.

A. Rubinstein. *Modeling Bounded Rationality.* MIT Press, 1998.

L.J. Savage. *The Foundations of Statistics.* Dover Publications, Inc., New York, 1954.

E. Schwitzgebel. The stanford encyclopedia of philosophy, August 2006. URL `http://plato.stanford.edu/`. E. Zalta (ed.).

J.R. Searle. *The Construction of Social Reality.* Allen Lane, London, 1995.

J.R. Searle. *Intentionality.* Cambridge University Press, 1983.

A. Sen. Why exactly is commitment important for rationality? *Economics and Philosophy*, 21(01):5–14, 2005.

A. Sen. *Rationality and Freedom.* Harvard University Press, Cambridge, MA, 2002.

A. Sen. *Collective Choice and Social Welfare.* Holden-Day, 1970.

R. Serrano. Fifty years of the nash program, 1953-2003. *Investigaciones Económicas*, XXIX(2), 2005.

K. Setiya. *Reasons without Rationalism.* Princeton University Press, 2007.

K. Setiya. Cognitivism about instrumental reason. *Ethics*, forthcoming.

N. Shah and J.D. Velleman. Doxastic deliberation. *The Philosophical Review*, forthcoming.

H.A. Simon. *Models of Bounded Rationality*, volume 1-2. MIT Press, 1982.

R. Stalnaker. *Inquiry.* MIT Press, 1984.

R. Sugden. The logic of team reasoning. *Philosophical Explanations 6*, pages 165–181, 2003.

B. ten Cate. *Model Theory for Extended Modal Languages*. PhD thesis, Universiteit van Amsterdam, ILLC Dissertation Series DS-2005-01, 2005.

R. Tuomela. *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford University Press, Stanford, 1995.

J. van Benthem. Logic in games, electronic lecture notes, www.illc.uva.nl/lgc, 1999.

J. van Benthem. Extensive games as process models. *Journal of Logic, Language and Information*, 11:289–313, 2002.

J. van Benthem. Rational dynamic and epistemic logic in games. In S. Vannucci, editor, *Logic, Game Theory and Social Choice III*, pages 19–23. University of Siena, department of political economy, 2003. An updated version of this paper is now available on `http://staff.science.uva.nl/~johan/RatDyn.2006.pdf`. The page numbering comes from this version.

J. van Benthem. Games in dynamic epistemic logic. *Bulletin of Economic Research*, 53(4):219–248, 2001.

J. van Benthem. Open problems in logic and games. In S.N. Artemov, H. Barringer, A.S. d'Avila Garcez, L.C. Lamb, and J. Woods, editors, *We Will Show Them! Essays in Honour of Dov Gabbay*, volume 1. College Publications, 2005.

J. van Benthem. Epistemic logic and epistemology, the state of their affairs. *Philosophical Studies*, 128:49 – 76, 2006.

J. van Benthem. One is a lonely number. In Z. Chatzidakis, P. Koepke, and W. Pohlers, editors, *Logic Colloquium '02*, page 96 129, Wellesley MA, 2006a. ASL & A.K. Peters.

J. van Benthem. Open problems in logical dynamics. In D. Gabbay, S. Goncharov, and M. Zakharysashev, editors, *Mathematical Problems from Applied Logic I*, page 137 192, New York & Novosibirsk, 2006b. Springer.

J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-classical Logics*, 17(2), 2007.

J van Benthem. Characterizing update rules. Universiteit van Amsterdam, Manuscript.

J. van Benthem and F. Liu. Diversity of agents in games. *Phiosophia Scientiae*, 8(2), 2004.

J. van Benthem, B. Kooi, and J. van Eijck. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2005.

J. van Benthem, S. van Otterloo, and O. Roy. Preference logic, conditionals, and solution concepts in games. In *Festschrift for Krister Segerberg*. University of Uppsala, 2005.

J. van Benthem, P. Girard, and O. Roy. Everything else being equal: A modal logic approach to ceteris paribus preferences. *Journal of Philosophical Logic*, Forthcoming.

W. van der Hoek, W. Jamroga, and M. Wooldrige. Towards a theory of intention revision. *Synthese*, 155, March 2007. Knowledge, Rationality & Action 103-128.

H. van Ditmarsch, W. van de Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library Series*. Springer, 2007.

M. van Hees and O. Roy. Intentions and plans in decision and game theory. In B. Verbeek, editor, *Reasons and Intentions*. Ashgate Publishers, 2007a.

M. van Hees and O. Roy. Intentions, decisions and rationality. In T. Boylan and R. Gekker, editors, *Economics, Rational Choice and Normative Philosophy*. Routhledge, May 2007b.

S. van Otterloo. *A Strategic Analysis of Multi-Agent Protocols*. PhD thesis, University of Liverpool, 2005.

J.D. Velleman. *Self to Self*. Cambridge University Press, 2005.

J.D. Velleman. What good is a will? Downloaded from the author's website on April 5th 2006, April 2003.

J.D. Velleman. How to share an intention. *Philosophy and Phenomenological Research*, 57(1):29–50, Mars 1997.

B. Verbeek. *Moral Philosophy and Instrumental Rationality: an Essay on the Virtues of Cooperation*. Kluwer Academic Publishers, 2002.

J. von Neumann and O. Morgenstern. *A Theory of Games and Economic Behaviour*. Princeton University Press: Princeton, NJ, 1944.

G. von Wright. *The logic of preference*. Edinburgh University Press, 1963.

R.J. Wallace. Normativity, commitment, and instrumental reason. *Philosophers' Imprint*, 1(3):1–26, 2003a.

R.J. Wallace. *Normativity and the Will*. Oxford University Press, 2006.

R.J. Wallace. The stanford encyclopedia of philosophy, October 2003b. URL `http://plato.stanford.edu/`. E. Zalta (ed.).

B. Willams. Ethical consistency. In *Problems of the Self*. Cambridge University Press, 1973.

L. Wittgenstein. *Philosophical Investigations*. Blackwell Publishers, 2001. Translated from German by G. E. M. Anscombe.

M. Wooldridge. *Reasoning about Rational Agents*. Intelligent robotics and autonomous agents series. MIT Press, Cambirdge, 2000.

G. Yaffe. Trying, intending, and attempted crimes. *Philosophical Topics*, 32(1-2), 2004.

Index

Samenvatting

In dit proefschrift ontwikkel ik een theorie van beslissingen die recht doet aan het feit dat mensen toekomstgerichte intenties vormen. Tegelijkertijd maak ik gebruik van moderne theorieën van instrumentele rationaliteit en dynamisch epistemische logica. Het resultaat is een completer beeld van praktisch redeneren. Ik laat zien dat een zodanige benadering van het vraagstuk bestaande theorieën van rationeel beslissen en intenties verrijkt.

Hoofdstuk 2 laat zien dat de introductie van toekomstgerichte intenties inderdaad de verklarende kracht van beslis-theoretische modellen versterkt. Met toekomstgerichte intenties verbinden agenten zich aan een bepaalde uitkomst en dit staat ons toe om genuanceerder om te gaan met op het eerste gezicht evenredig attractieve uitkomsten. Tegelijkertijd verklaren intenties beter hoe agenten hun beslissingen over tijd coördineren. Dit brengt de traditionele beslistheorie een stap verder.

Hoofdstuk 3 bespreekt coördinatie tussen verschillende agenten, vooral in “Hi-Lo games”. Ik laat zien dat intenties inderdaad helpen om coördinatie—ook tussen verschillende agenten—beter worden verankerd, op een manier die generaliseert van één naar meerdere agenten. Aan het eind van het hoofdstuk laat ik zien hoe intenties in het algemeen (niet alleen in “Hi-Lo games”) coördinatie verankeren. Dit staat ons toe om belangrijke beweringen met betrekking tot gemeenschappelijk handelen in de filosofie te verklaren.

In hoofdstuk 4 bespreek ik twee facetten van het bindende vermogen van intenties en hun invloed op praktisch redeneren: Eerst het filteren van mogelijkheden en vervolgens de focus op middelen. Ik laat zien dat beide onderwerpen kunnen worden verklaard als transformaties van beslis- en speltheoretische modellen. In de context van strategische interactie krijgen deze onderwerpen een belangrijk sociaal karakter dat nog niet eerder bestudeerd is in de filosofische theorie van actie.

In hoofdstuk 5 maak ik gebruik van dynamisch epistemische logica om de ideeën uit voorgaande hoofdstukken tot één theorie te integreren. Ik laat be-

langrijke overeenkomsten zien tussen de rol van intenties in coördinatie en het filteren van mogelijkheden. Deze observatie leidt tot een natuurlijke generalisatie van het filterproces die rekening houdt met de informatie die agenten tot hun beschikking hebben over hun eigen en andermans intenties. Vervolgens bespreek ik hoe onder andere het filteren en de focus op middelen helpen coördinatie te verklaren en hoe ze benvloed worden door de bekende oplossingsconcepten.

In hoofdstuk 6 neem ik een meer filosofisch perspectief en ik bespreek hoe de normen van consistentie en coherentie van intenties kunnen worden verklaard. Er bestaan twee dominante verklaringen in de hedendaagse filosofische theorie van actie: de "cognitieve" en de "agency" benadering. Ik ontwikkel een alternatieve benadering die omschreven kan worden als hybride pragmatisme en tussen de twee andere concepten geplaatst kan worden. Hybride pragmatisme is gebaseerd op het tot nu toe weinig besproken concept van "acceptance in deliberation". Ik beargumenteer dat hybride pragmatisme beter kan verklaren hoe toekomstgerichte intenties invloed hebben op praktisch redeneren.

Résumé

Cette thèse développe une théorie de la prise de décision qui, tout en s'appuyant sur les théories contemporaines de la rationalité instrumentale et sur la logique épistémique dynamique, rend justice au fait que les agents humains ont la capacité de former des intentions à propos d'actions futures. Il en résulte un point de vue plus complet sur le raisonnement pratique, qui enrichit à la fois les modèles existants de la prise de décision rationnelle et les théories philosophiques des intentions.

Dans le chapitre 2 je montre qu'en introduisant des intentions à propos d'action futures on élargit le pouvoir explicatif des modèles en théorie de la décision. L'engagement volitif associé aux intentions, par exemple, permets aux agents de départager des options qui ne peuvent être distinguées dans les modèles classiques, et par le fait même de mieux coordonner leur actions dans le temps.

Dans le chapitre 3 j'étudie à nouveau la coordination, mais cette fois en contextes interactifs, plus particulièrement dans les jeux de type *Hi-Lo*. Je montre que les intentions facilitent la coordinations entre agents dans ces contextes spécifiques, d'une manière qui généralise naturellement les résultats obtenus au chapitre précédent. J'examine ensuite comment les intentions facilitent la coordination, non seulement dans les jeux de type Hi-Lo, mais de manière plus générale. Ceci permet de jeter un oeil nouveau sur des hypothèses importantes en philosophie de l'action, notamment sur la notion d'action conjointe.

Le chapitre 4 se consacre à deux facettes de l'engagement généré par les intentions dans le raisonnement pratique: la pré-sélection d'options et le ciblage de la délibération sur les moyens de parvenir aux fins. Je montre qu'elles peuvent être étudiées sous la forme de deux opérations simples qui transforment les modèles en théorie de la décision et des jeux. En situations d'interactions, ces opérations prennent une dimension sociale importante, qui n'a précédemment pas été étudiée en philosophie de l'action.

Dans le chapitre 5 j'utilise la logique épistémique dynamique pour construire une théorie du raisonnement pratique englobant et généralisant les résultats

obtenus dans les chapitres précédents. Je montre qu'un aspect important de l'engagement volitif utilisé au Chapitre 3 dans l'étude de la coordination interpersonnelle réapparaît dans la pré-sélection d'options définie au Chapitre 4. De cette observation découle une généralisation naturelle du concept de pré-sélection, qui prend en compte l'information que les agents possèdent à propos de leurs propres intentions et de celles des autres agents. En fin de chapitre j'explore deux autres thèmes à l'intersection des théories de l'action et de la rationalité instrumentale, soit les conditions sous lesquelles la pré-sélection d'options facilite la coordination et est à son tour facilitée par l'élimination d'options irrationnelles.

Le chapitre 6 propose un retour philosophique sur la théorie développée au chapitre précédent, pour mettre en lumière la provenance de différentes normes associés aux intentions. Deux approches, le cognitivisme et l'approche pratique, dominent le débat sur cette question en philosophie contemporaine de l'action. Dans ce chapitre je développe une approche alternative, que je nomme le pragmatisme hybride, située à mi-chemin entre le cognitivisme et l'approche pratique. Le pragmatisme hybride se base sur le concept d'acceptation en délibération, un état cognitif qui n'a jusqu'à maintenant que peu été étudié en philosophie. Je soutiens que le pragmatisme hybride constitue une alternative avantageuse, et que son recours aux acceptations en délibération permet de mieux comprendre comment les intentions à propos d'action futures influencent le raisonnement pratique.

Abstract

In this thesis I propose a theory of decision making that does justice to the idea that human agents can form future-directed intentions, but which at the same time capitalizes on the resources of contemporary theories of instrumental rationality and dynamic epistemic logic. The result is a more all-encompassing picture of practical reasoning for planning agents. I show that such a broad approach genuinely enriches existing models of rational decision making, as well as the philosophical theory of intentions.

In Chapter 2 I show that the introduction of future-directed intentions does indeed broaden the explanatory scope of *decision*-theoretic models. The volitive commitment of future-directed intentions allows one to go beyond traditional decision-theoretic reasoning by "breaking ties" between equally desirable options, and thus provides a straightforward anchor for personal coordination.

In Chapter 3 I consider coordination, mostly in "Hi-Lo" games. I show that intentions do indeed anchor coordination in these games, in a way that naturally generalizes their "tie-breaking" effect in single agent contexts. At the end of the chapter I look at how intentions can anchor coordination in the general case. This allows to revisit important claims in the planning theory concerning "shared agency", and in particular to circumscribe better the extent of this phenomenon.

In Chapter 4 I turn to two facets of the reasoning-centered commitment of intentions, namely the filtering of options and the focus on means. I show that they can be studied by means of two simple operations which transform decision- and game-theoretic models. In contexts of strategic interaction, these operations acquire an important social character, that has not yet been studied in philosophy of action.

In Chapter 5 I use dynamic epistemic logic to bring the considerations of the previous chapters under a single umbrella. I show that an important aspect of the volitive commitment used to account for coordination with intentions has an echo in the filtering of options that I define in Chapter 4. This observation triggers a natural generalization of the idea of filtering, which takes into account the information that agents have about their own intentions and the intentions of others. By the end of the chapter I explore two other issues at the intersection of planning agency and instrumental rationality, namely the condition under which intention-based transformations of decision problems foster coordination and become “enabled” by the elimination of dominated strategies.

In Chapter 6 I look back at this theory from a philosophical point of view, and investigate the question of how the norms of consistency and coherence which apply to intentions can be explained. In contemporary philosophy of action there are two main takes on this issue, called the “cognitivist” and “agency” approaches. Here I explore an alternative one, *hybrid pragmatism*, which stands half-way between cognitivism and the agency approach. It is based on the notion of “acceptance in deliberation”, a cognitive state which has so far attracted little attention. I argue that hybrid pragmatism is a plausible alternative to the two main contemporary approaches, and that its use of acceptances provides a more complete picture of how future-directed intentions make their way into practical reasoning.

Titles in the ILLC Dissertation Series:

ILLC DS-2001-01: **Maria Aloni**
Quantification under Conceptual Covers

ILLC DS-2001-02: **Alexander van den Bosch**
Rationality in Discovery - a study of Logic, Cognition, Computation and Neuropharmacology

ILLC DS-2001-03: **Erik de Haas**
Logics For OO Information Systems: a Semantic Study of Object Orientation from a Categorial Substructural Perspective

ILLC DS-2001-04: **Rosalie Iemhoff**
Provability Logic and Admissible Rules

ILLC DS-2001-05: **Eva Hoogland**
Definability and Interpolation: Model-theoretic investigations

ILLC DS-2001-06: **Ronald de Wolf**
Quantum Computing and Communication Complexity

ILLC DS-2001-07: **Katsumi Sasaki**
Logics and Provability

ILLC DS-2001-08: **Allard Tamminga**
Belief Dynamics. (Epistemo)logical Investigations

ILLC DS-2001-09: **Gwen Kerdiles**
Saying It with Pictures: a Logical Landscape of Conceptual Graphs

ILLC DS-2001-10: **Marc Pauly**
Logic for Social Software

ILLC DS-2002-01: **Nikos Massios**
Decision-Theoretic Robotic Surveillance

ILLC DS-2002-02: **Marco Aiello**
Spatial Reasoning: Theory and Practice

ILLC DS-2002-03: **Yuri Engelhardt**
The Language of Graphics

ILLC DS-2002-04: **Willem Klaas van Dam**
On Quantum Computation Theory

ILLC DS-2002-05: **Rosella Gennari**
Mapping Inferences: Constraint Propagation and Diamond Satisfaction

ILLC DS-2002-06: **Ivar Vermeulen**
A Logical Approach to Competition in Industries

ILLC DS-2003-01: **Barteld Kooi**
Knowledge, chance, and change

ILLC DS-2003-02: **Elisabeth Catherine Brouwer**
Imagining Metaphors: Cognitive Representation in Interpretation and Understanding

ILLC DS-2003-03: **Juan Heguiabehere**
Building Logic Toolboxes

ILLC DS-2003-04: **Christof Monz**
From Document Retrieval to Question Answering

ILLC DS-2004-01: **Hein Philipp Röhrig**
Quantum Query Complexity and Distributed Computing

ILLC DS-2004-02: **Sebastian Brand**
Rule-based Constraint Propagation: Theory and Applications

ILLC DS-2004-03: **Boudewijn de Bruin**
Explaining Games. On the Logic of Game Theoretic Explanations

ILLC DS-2005-01: **Balder David ten Cate**
Model theory for extended modal languages

ILLC DS-2005-02: **Willem-Jan van Hoeve**
Operations Research Techniques in Constraint Programming

ILLC DS-2005-03: **Rosja Mastop**
What can you do? Imperative mood in Semantic Theory

ILLC DS-2005-04: **Anna Pilatova**
A User's Guide to Proper names: Their Pragmatics and Semanics

ILLC DS-2005-05: **Sieuwert van Otterloo**
A Strategic Analysis of Multi-agent Protocols

ILLC DS-2006-01: **Troy Lee**
Kolmogorov complexity and formula size lower bounds

ILLC DS-2006-02: **Nick Bezhanishvili**
Lattices of intermediate and cylindric modal logics

ILLC DS-2006-03: **Clemens Kupke**
Finitary coalgebraic logics

ILLC DS-2006-04: **Robert Špalek**
Quantum Algorithms, Lower Bounds, and Time-Space Tradeoffs

ILLC DS-2006-05: **Aline Honingh**
The Origin and Well-Formedness of Tonal Pitch Structures

ILLC DS-2006-06: **Merlijn Sevenster**
Branches of imperfect information: logic, games, and computation

ILLC DS-2006-07: **Marie Nilsenova**
Rises and Falls. Studies in the Semantics and Pragmatics of Intonation

ILLC DS-2006-08: **Darko Sarenac**
Products of Topological Modal Logics

ILLC DS-2007-01: **Rudi Cilibrasi**
Statistical Inference Through Data Compression

ILLC DS-2007-02: **Neta Spiro**
What contributes to the perception of musical phrases in western classical music?

ILLC DS-2007-03: **Darrin Hindsill**
It's a Process and an Event: Perspectives in Event Semantics

ILLC DS-2007-04: **Katrin Schulz**
Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals

ILLC DS-2007-05: **Yoav Seginer**
Learning Syntactic Structure

ILLC DS-2008-01: **Stephanie Wehner**
Cryptography in a Quantum World

ILLC DS-2008-02: **Fenrong Liu**
Changing for the Better: Preference Dynamics and Agent Diversity

ILLC DS-2008-03: **Olivier Roy**
Thinking before Acting: Intentions, Logic, Rational Choice