Thinking before acting: intentions, logic, rational choice
Roy, O.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
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).
Chapter 3. Intentions and coordination in strategic games

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 \(G\) is a tuple \(⟨I,S,\pi,\preceq⟩⟩\) such that:

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

\(^2\)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 \(S\), for one agent \(i\). A mixed strategy is a probability distribution on \(S\).
3.1. Intentions in strategic interactions

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

The definition of the outcome function $\pi$ 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.

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.

<table>
<thead>
<tr>
<th></th>
<th>Cinema</th>
<th>Restaurant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cinema</td>
<td>together</td>
<td>alone</td>
</tr>
<tr>
<td>Restaurant</td>
<td>alone</td>
<td>together</td>
</tr>
</tbody>
</table>

Table 3.1: A coordination game.

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].
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. 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 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
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 Def-
nition 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 coor-
dination in strategic games, starting with the “easy” case of Hi-Lo games and
then generalizing to arbitrary strategic contexts. Along the way I introduce epis-
temic 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 \( G \), 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 \( \iota_i \)
is said to be generated by \( \downarrow \iota_i \). An intention profile \( \iota \) 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 coordi-
nation. 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 ac-
count 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.”
Chapter 3. Intentions and coordination in strategic games

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 $G$ such that:

- $S_i = S_j$ for all $i, j \in I$
- $\pi(\sigma) \succ_i \pi(\sigma')$ for all $\sigma$ such that $\sigma(i) = \sigma(j)$ for all $i, j \in I$ and $\sigma'$ 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 $G$, a strategy profile $\sigma$ 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 $\sigma$ 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 $\sigma$ 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.
3.2. Coordination in Hi-Lo games

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 $\downarrow \iota$ has to contain only
3.3. Intentions and rational expectations

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

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0, 0</td>
<td>0, 7</td>
</tr>
<tr>
<td>b</td>
<td>7, 0</td>
<td>1, 1</td>
</tr>
</tbody>
</table>

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 \(\Gamma\) 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 \(\Gamma\), or \(\Gamma\)-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 \(\Gamma\)-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).
Chapter 3. Intentions and coordination in strategic games

An intention set \( \iota_i \) is said to be \textit{payoff-compatible} whenever \( A \in \iota_i \) if \( A \cup B \in \iota_i \), \( A^* \neq \emptyset \) and \( x \succeq y \) 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 \( \Gamma \) 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 \( \Gamma \) be a solution concept and \( \Gamma_{\succeq, i} \) be the set of most preferred \( \Gamma \)-feasible outcomes for agent \( i \). For all \( i \) and \( \iota_i \) as in Definition 3.1.2, \( \iota_i \) is payoff-compatible if and only if there is a non-empty \( A \subseteq \Gamma_{\succeq, i} \) such that \( \iota_i = \{ B : A \subseteq B \} \).

\textbf{Proof.} Essentially the same as for Fact 2.2.9, adapted to feasible sets. \( \blacksquare \)

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 \( \sigma^* \) 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 \textit{intend} to realize the outcome of this profile, and not why they would \textit{actually} coordinate. To coordinate successfully on the basis of such overlapping intentions they still need to \textit{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 \textit{we} achieve the Pareto-optimal solution.” As I mentioned at the end of Section 3.1, there are arguably cases where the agents are \textit{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 \(G\) 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* \(M\) of the game \(G\) is a tuple \(\langle W, f, \{\sim_i\}_{i \in I}\rangle\) such that:

- \(W\) is a set of states.
- \(f : W \rightarrow \Pi_{i \in I} S_i \times \Pi_{i \in I} F(X)\) is a function that assigns to each \(w \in W\) a pair \((\sigma, \iota)\) 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.
Chapter 3. Intentions and coordination in strategic games

and \(\iota(w)\) for the \(\sigma\) (alternatively the \(\iota\)) 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 \(M, w\) is an epistemic model for the game \(G\) together with a distinguished state \(w\), the actual play of \(G\).

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 \(K_i(W - K_i E) \subseteq W - K_i E\), where \(W - A\) is the complement of \(A\) in \(W\).
3.4. Epistemic models for games with intentions

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 ($\iota_i(w) = \iota_i(w')$)\footnote{These conditions are illustrated in Figures 5.1 and 5.2, on page 94 and 95.}. 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 that 2 might play Hi as well as Lo. But 1 knows what he plays at 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.](image-url)
$\iota_1(Hi - Lo)$ 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 $M, w$ 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 $M, w \}$

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. Coordination with payoff-compatible intentions

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

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

2. If $\sigma^*$ is the Pareto-optimal strategy profile of that game, then we can construct a pointed model $M, w$ 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 $\sigma'$ 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 $\sigma'$ such that $\pi(\sigma') = x$. This means that $\sigma'$ can only be $\sigma^*$, 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 $[w]$, 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.
Chapter 3. Intentions and coordination in strategic games

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 \( G \) the following are equivalent.

1. There is a weakly Pareto-optimal profile \( \sigma^* \).

2. There is an epistemic pointed model \( M, w \) for \( G \) such that at \( w \) all agents have payoff-compatible intentions, taking \( \Gamma = \Pi_{i \in I} S_i \), and \( \bigcap_i \downarrow \mu_i \neq \emptyset \).

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

From (1) to (2), take any such \( \pi(\sigma^*) \). Any pointed model \( M, w \) built according to Definition 5.2.1 in which \( \downarrow \mu_i = \{ \pi(\sigma^*) \} \) 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, a 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 \( w'' \sim_j w' \).
Chapter 3. Intentions and coordination in strategic games

<table>
<thead>
<tr>
<th>State</th>
<th>$\downarrow_{1}(w)$</th>
<th>$\downarrow_{2}(w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Hi - Hi$</td>
<td>$Hi - Hi$</td>
<td>$Hi - Hi$</td>
</tr>
<tr>
<td>$Hi - Lo$</td>
<td>$Hi - Hi$</td>
<td>$Hi - Hi$</td>
</tr>
<tr>
<td>$Lo - Hi$</td>
<td>$Lo - Lo$</td>
<td>$Hi - Hi$</td>
</tr>
<tr>
<td>$Lo - Lo$</td>
<td>$Lo - Lo$</td>
<td>$Lo - Lo$</td>
</tr>
</tbody>
</table>

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 $\lambda$, 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 $M, w$, if all agents have strongly epistemically supported intentions to achieve $\{\pi(\sigma^*)\}$, then $\sigma(w)$ is the Pareto-optimal strategy profile $\sigma^*$ of that game.

2. If $\sigma^*$ is the Pareto-optimal profile of that game, then we can construct a pointed model $M, w$ 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 $\nu_i(w)$ is payoff-compatible, and $\Gamma$ be the set of pure Nash equilibria. Then for any Hi-Lo game the following holds:

1. For any of its pointed model $M, w$, 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 $\sigma^*$ of that game.
2. If \( \sigma^* \) is the Pareto-optimal profile of that game, then we can construct a pointed model \( \mathcal{M},w \) 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 \( \mathcal{K}_i(E) \subseteq E \) 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 \( \mathcal{M},w \) 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 has 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 \( G \) 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.
3.6. Stackelberg heuristic and intention-based coordination

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

- 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 $s^h_i$ be a strategy of $i$ that yields a most preferred Stackelberg outcome. A two-agent strategic game $G$ is Stackelberg soluble if there is a $\sigma^h$ such that $\sigma^h(i) = s^h_i$ for all $i \in I$. $\sigma^h$ 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 $s^h_i$. 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, $\sigma$ 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 $G$ with intentions in which $s^h_i$ is unique, $i$ is payoff-compatible for both agents, and $\Gamma$ is the set of pure Nash equilibria.

- $G$ is Stackelberg soluble with $\sigma^h$.

- $\bigcap_{i \in I} i = \{\pi(\sigma^h)\}$. 
Chapter 3. Intentions and coordination in strategic games

Proof. \( G \) is Stackelberg soluble with \( \sigma^h \) iff \( \sigma^h \) is a Nash equilibrium, thus iff \( \pi(\sigma^h) \) is a \( \Gamma \)-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 \( \pi(\sigma') \geq_i \pi(\sigma^h) \). Indeed, if there were such a \( \sigma' \), 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 \( s_i^h \) is the unique strategy that yields the most preferred Stackelberg outcome, and similarly for player 2. This means that \( G \) is Stackelberg soluble with \( \sigma^h \) 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 \( \nu_i = \{\pi(\sigma^h)\} \) for both players. \( \blacksquare \)

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.
3.8. A general account of intention-based coordination

<table>
<thead>
<tr>
<th></th>
<th>Boxing</th>
<th>Ballet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boxing</td>
<td>(2,1)</td>
<td>(0,0)</td>
</tr>
<tr>
<td>Ballet</td>
<td>(0,0)</td>
<td>(1,2)</td>
</tr>
</tbody>
</table>

Table 3.5: The Battle of the Sexes.

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

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 \( \sigma(w) = \text{Boxing} - \text{Ballet} \).

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:  

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 \( M \) 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 \( G \), 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 \( G \) and a pointed model \( M, w \) 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 \( \sigma_1 \) 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 \( G \) and a pointed model \( M, w \) 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.
3.8. A general account of intention-based coordination

![Diagram of a model for independence proofs](image)

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 $\downarrow_{i}(\sigma_1) = W$ and $\downarrow_{i}(\sigma_1) = A$ for the other. We get that at $\sigma_1$ 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 $G$ and a pointed model $M, w$ 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 $\downarrow_{i}(\sigma_1) = \{\sigma_4\}$ and fix $\downarrow_{i}(\sigma_1) = \{\sigma_1\}$ 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 $G$ be a game let $C \subset \Pi_{i \in I} S_i$ be the non-empty set of coordination profiles.

- For any epistemic model $M$ for $G$ 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 $\sigma \not\in C$ then we can construct a model $M$ 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 $G$ and a pointed model $M, w$ 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. 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.

---

23Once 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 $G$, 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 $SD(G)$ which results after elimination of strictly dominated strategies from $G$ is defined as follows:

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

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

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 $\sigma$ 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].

\[\text{Removal of weakly dominated strategies is more tricky to define. See Myerson [1991, p.90].}\]