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.
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 \(\mathcal{F}\) 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 \(w' \nprec w\).

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. Preliminaries: modal logic for preferences

5.1.2. Definition. [Preference language] Given a set of atomic proposition $PROP$, the language $L_P$ is inductively defined as follows:\footnote{A few guidelines for the reader unaccustomed to this way of defining logical languages. $\phi ::= p \mid \phi \land \psi \mid \neg \phi \mid \Diamond \leq \phi \mid \Diamond < \phi \mid E\phi$ means that a formula $\phi$ 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.}

$$\phi ::= p \mid \phi \land \psi \mid \neg \phi \mid \Diamond \leq \phi \mid \Diamond < \phi \mid E\phi$$

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 “$\phi$ is true in a state that is considered at least as good as the current state” and “$\phi$ 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 $\phi$ 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 “$\phi$ holds in all states that are at least as good as the current one”. $\Box <$ is defined similarly. $A\phi$, which abbreviates $\neg E \neg \phi$, is taken to mean “$\phi$ 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 $M$ is a preference frame $F$ 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 $M, w$.

5.1.4. Definition. [Truth and validity in $L_P$]

- $M, w \models p \iff w \in V(p)$
- $M, w \models \phi \land \psi \iff M, w \models \phi$ and $M, w \models \psi$
- $M, w \models \neg \phi \iff M, w \not\models \phi$
- $M, w \models \Diamond \leq \phi \iff$ there is a $v$ such that $v \geq w$ and $M, v \models \phi$
- $M, w \models \Diamond < \phi \iff$ there is a $v$ such that $v > w$ and $M, v \models \phi$
- $M, w \models E\phi \iff$ there is a $v$ such that $M, v \models \phi$

A formula $\phi$ is valid in a preference model $M$, denoted $M \models \phi$, whenever $M, w \models \phi$ for all $w \in W$. A formula is valid in a preference frame $F$ whenever it is valid in
all preference models $\mathcal{M} = \langle \mathcal{F}, \mathcal{V} \rangle$. Finally, a formula is valid in a class of models $\mathcal{M}$ whenever it is valid in all models $\mathcal{M} \in \mathcal{M}$. 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 “$\phi$ 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 $\mathcal{L}_P$, 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 $\mathcal{L}_P$ 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 $\phi$ such that $\phi$ is valid in a class of frame $\mathcal{F}$ iff all the frames in $\mathcal{F}$ 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 $\mathcal{L}_P^3$:

\[
\begin{align*}
\Diamond \leq \Diamond \leq \phi & \rightarrow \Diamond \leq \phi \quad \text{(Transitivity)} \\
\phi & \rightarrow \Diamond \leq \phi \quad \text{(Reflexivity)}
\end{align*}
\]

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.

\[
\phi \land E\psi \rightarrow (\Diamond \leq \psi \lor E(\psi \land \Diamond \leq \phi)) \quad \text{(Totality)}
\]

\[
^3\text{The correspondence arguments are well know. See again Blackburn et al. [2001, chap.3].}
\]
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 $M$ where this formula is valid, which contains only two states $w$ and $w'$ and where $\phi$ and $\psi$ are true only at $w$ and $w'$, respectively. The truth condition for $E$ gives $M, w \models \phi \land E\psi$, and so the consequent of (Totality) must also be true there. But then either $M, w \models \Diamond \preceq \psi$, which means that $w' \succeq w$ or $M, w \models E(\psi \land \Diamond \preceq \phi)$, which means, by the way we devised our model, that $w \succeq w'$.

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

$$\Diamond \prec \phi \rightarrow \Diamond \preceq \phi$$  \hfill (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 $L_P$. That is, there is no formula of $L_P$ 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 $M, w$ and $M', v$ 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 $M'$ to $M$. A standard modal logic argument shows that if two pointed preference models are bisimilar, then they make exactly the same formulas of $L_P$ true.

With this in hand, to show that a given property is not definable in $L_P$ 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 $L_P$. 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 $L_P$. 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 $\Diamond \prec$ is not definable in terms of $\Diamond \preceq$. In other words, to talk directly about the $\succ$ relation one has to introduce a separate modality.
Chapter 5. Logics for practical reasoning with intentions

\[ \phi \leq_{\forall \exists} \psi \iff A(\phi \rightarrow \Diamond \leq \psi) \quad \text{(Lifted relation)} \]

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

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 \( \phi \) and \( \psi \) of \( L_P \), either \( \phi \leq_{\forall \exists} \psi \) or \( \psi \leq_{\forall \exists} \phi \).

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

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. 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 $\Lambda^{L^P}$ 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 $G$. 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 $G$ 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) $\Box \leq (\phi \land \psi) \leftrightarrow \Box \leq \phi \land \Box \leq \psi$
  (Trans) $\Diamond \leq \Box \leq \phi \rightarrow \Diamond \leq \phi$
  (Ref) $\phi \rightarrow \Diamond \leq \phi$
  (Tot) $\phi \land E\psi \rightarrow (\Diamond \leq \psi \lor E(\psi \land \Diamond \leq \phi))$

• For $\Diamond <$:
  (K) $\Box < (\phi \land \psi) \leftrightarrow \Box < \phi \land \Box \leq \psi$

• S5 for $E$:
  (K) $A(\phi \land \psi) \leftrightarrow A\phi \land A\psi$
  (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$_2$ $\phi \land \Diamond \leq \psi \rightarrow (\Diamond < \psi \lor \Diamond \leq (\psi \land \Diamond \leq \phi))$
  Int$_3$ $\Diamond < \Diamond \leq \phi \rightarrow \Diamond < \phi$

• The following inference rules:

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

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

5.2.1. Definition. [Epistemic game frames with intentions] An epistemic game frame with intentions $G$ is a tuple $\langle I, W, \{i, \sim_i, \succeq_i\}_{i \in I} \rangle$ 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 $\{w' : w \sim_i w'\}$. 


5.2. Logic for games with intentions

• \( \succeq_i \) 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 \( i(w) = i(w') \).

The intention function \( i \) 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”—\( \sigma, \sigma' \) 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 \( \text{PROP} \) and a set of nominals \( S \), let \( L_{GF} \) be the language defined as:

\[
\phi ::= p \mid \sigma \mid \phi \land \phi \mid \neg \phi \mid \Diamond \leq \phi \mid \Diamond < \phi \mid K_i \phi \mid I_i \phi \mid E \phi
\]

We are now familiar with the preference fragment of that language. Formulas of the form \( K_i \phi \) should be read “\( i \) knows that \( \phi \)” and those of the form \( I_i \phi \) as “\( i \) intends that \( \phi \)” As in the previous section, these connectives have duals. For \( \neg K_i \neg \phi \) I use \( \Diamond i \phi \), which means “\( i \) considers \( \phi \) possible”, and for \( \neg I_i \neg \phi \) I use \( i \phi \), meaning “\( \phi \) 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* \( M \) is an epistemic game frame \( G \) 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 \( M, w \).

\(^8\)Key references on hybrid logic are Blackburn et al. [2001, chap.7] and ten Cate [2005].
5.2.4. Definition. [Truth in $L_{GF}$] Formulas of the form $\Diamond_i \preceq \phi$, $\Diamond_i \preceq_1 \phi$ and $E \phi$ are interpreted as in 5.2.4.

\[
\begin{align*}
M, w \models \sigma & \quad \text{iff} \quad w \in V(\sigma). \\
M, w \models K_i \phi & \quad \text{for all } w' \text{ such that } w \sim_i w', M, w' \models \phi. \\
M, w \models I_i \phi & \quad \text{iff } ||\phi|| \in \iota_i(w), \text{ where } ||\phi|| = \{w' : M, w' \models \phi\}.
\end{align*}
\]

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 $\iota_i$ is a neighbourhood function. $I_i \phi$ is true at a state $w$ if and only if $i$ intends that $\phi$ at that state, i.e. if and only if the interpretation of $\phi$ is in the intention set of $i$ at $w$.

Here the assumption that $\iota_i$ 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 $\phi$ at $w$” if and only if $||\phi|| \in \iota_i(w)$, one can just say that “agent $i$ intends to achieve $\phi$ at $w$” if and only if $\phi$ 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 $\phi$ at $w$ if and only if $\phi$ 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 $\iota_i$, 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 $\iota_i$ 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 $i_\phi$. In the general case, i.e. when the neighbourhood are not necessarily filters, one has to include a separate clause to ensure that $i_\phi$ 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 $\iota_i$ 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 “\( \phi \) is compatible with \( i \)’s intentions”. One then requires that \( M, w \models i_\phi \) iff \( W - ||\phi|| \not\in \iota_i(w) \). But given that \( \iota_i(w) \) is a filter, we automatically get that \( M, w \models i_\phi \) iff there is a \( w' \in \iota_i(w) \) such that \( M, w' \models \phi \). 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 \( i_\phi \) is true whenever \( \phi \) is compatible with \( i \)’s intentions.

I am now ready to put \( L_{GF} \) 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 \( L_{GF} \) 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 \( L_{GF} \). 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\(^\text{11} \).

\[
\begin{align*}
s_i &\iff \bigvee_{\sigma(i)=s_i} \sigma \\
s_i \rightarrow K_i s_i & \quad \text{(Knowledge of strategy choice)}
\end{align*}
\]

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 \( L_{GF} \). On the one hand, a standard argument in neighbourhood semantic shows that \( \iota_i \) is closed under conjunction and disjunction, so

\(^\text{11}\)In the following I slightly abuse notation and write \( \sigma(i) \) instead of \( V(\sigma)(i) \).
that it is a filter, whenever the following hold\textsuperscript{12}.

\[ I_i(\phi \land \psi) \leftrightarrow I_i\phi \land I_i\psi \]  
(Intention closure)

For consistency, we have the following.

\[ i_i \top \]  
(Intention consistency)

Here, once again, the fact that we can see $i_i \phi$ as true at $w$ whenever there is a $w' \in \downarrow \iota_i(w)$ such that $\phi$ holds at $w'$ makes the correspondence argument completely straightforward. Indeed, since $\top$ is true at all states, to ask for the validity of $i_i \top$ 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 $L_{GF}$ as follows.

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

\[ I_i\phi \rightarrow K_iI_i\phi \]  
(Knowledge of Intention)

\textbf{Proof.} The right-to-left direction is obvious. For left to right, take a model $M$ 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 $\phi$ true only at $w'$. By definition, we get that $M, w \models I_i\phi$, and thus that $M, w \models K_iI_i\phi$. This means that $M, w' \models I_i\phi$. But by the way we fixed the valuation of $\phi$, it has to be that $\downarrow \iota_i(w') = \{w'\}$, which means that $\iota_i(w) = \iota_i(w')$. \[ \blacksquare \]

\textsuperscript{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.
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 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.\footnote{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.}

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

\footnote{Recall the remarks about this in the Introduction, Section 1.2. I also revisit belief consistency in the next Chapter, Section 6.1.}
Chapter 5. Logics for practical reasoning with intentions

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.

\[ I_i \phi \rightarrow \Diamond_i \phi \]  

(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 \(w' \in \downarrow v_i(w)\). 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 \(F\) in a class \(F\). If \([w]_i \cap \downarrow v_i(w) \neq \emptyset\) for all \(w\), then \(F \models IK_i\).

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

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

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 \(L_{GF}\), as follows.

\[ \bigvee_{\sigma(i) = u(i)} i_i \sigma \]  

(Intentions-rationality \((IR_i)\) at \(w\))
5.2. Logic for games with intentions

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 $\mathcal{M}, w$, if $\mathcal{M}, w \models IK_i$ then $\mathcal{M}, w \models IR_i$. There are, however, models where the converse does not hold.

**Proof.** For the first part, I show the contrapositive. Assume that $\mathcal{M}, w \models \neg IR_i$. That is, $\mathcal{M}, w \models \bigwedge_{\sigma(i) = w(i)} \neg \iota_i \sigma$. 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 $w' \in \downarrow \iota_i(w)$, $w' \not\in [w]_i$. Take $s_i$ to be the collection of nominals $\sigma'$ such that $V(\sigma') \in \downarrow \iota_i(w)$. We get $\mathcal{M}, w \models I_i (\bigvee_{\sigma' \in s_i} \sigma') \land K_i \neg (\bigvee_{\sigma' \in s_i} \sigma')$.

For the second part, take any pointed model $\mathcal{M}, w$ where $[w]_i \subseteq \{w' : w'(i) = w(i)\}$, and fix $\downarrow \iota_i(w) = \{w' : w'(i) = w(i)\} - [w]_i$. We obviously get that
IK, w ⊩ IR_i. But by again using s^*_i as the collection of nominals σ' such that V(σ') ∈ |μ_i(w)|, we get that M, w ⊩ I_i(\bigvee_{σ' ∈ s^*_i} σ') ∧ K_i¬(\bigvee_{σ' ∈ s^*_i} σ') \blacksquare

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 ∼_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.

\[ \bigvee_{σ(i) = w(i)} i_σ ∧ □_i σ \] (Epistemic intentions-rationality (IR_i^*) 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 M, w; M, w ⊩ IK_i iff M, w ⊩ 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 M, w ⊩ I_iφ ∧ K_i¬φ for some φ. This means that ↓_{μ_i}(w) ⊆ ||φ|| and that ||φ|| ∩ [w]_i = ∅. We thus have ↓_{μ_i}(w) \cap [w]_i = ∅. But this means that for all σ' and w'(i) = w(i) such that V(σ') = w' and w' ∈ [w]_i, M, w ⊩ ¬i_σ'. In other words, M, w ⊩ \bigwedge_{σ'(i) = w(i)} □_i σ' → ¬i_σ'. \blacksquare

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) ≠ w_3(1), that at w_1 we have ↓_{μ_1}(w_1) = \{w_3\} and that PROP is empty. We clearly get M, w_1 ⊩ I_1φ ∧ □_1 φ while M, w_1 ̸⊩ IR_1 and M, w_1 ̸⊩ IR_i^*.

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, $M, w \models IR_i$ implies $M, w \models K_iIR_i$ and $M, w \models \neg IR_i$ implies $M, w \models K_i\neg IR_i$. 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 $M, w \models IR_i$. 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 $M, w'' \models IR_i$.

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\textsuperscript{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.

\textsuperscript{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:

\[ \phi >^i \forall \forall \psi \land I_i(\phi \lor \psi) \rightarrow I_i\phi \]  
( Payoff-compatibility (IPC$_i$) )

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

- $\mathcal{M}, w \models \bigwedge_{i \in I} K_i(\bigwedge_{j \neq i}(IPR_j \land IR_j))$
- $w$ 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: $\phi >^i \forall \forall \psi$. This formula states that $\phi$ is strictly preferred to $\psi$ whenever all states that make $\phi$ true are strictly preferred to all states that make $\psi$ 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 $K_i(\bigwedge_{j \neq i}(IPR_j \land IR_j))$ 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 $K_i(\bigwedge_{j \neq i}(IPR_j \land IR_j))$ can only be true at that profile. ■

For the first time in this section, this result explicitly uses the preferences fragment of $L^G\mathcal{F}$. 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 $L^G\mathcal{F}$ 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 $L^G\mathcal{F}$ and which “pools” the epistemic accessibility relations

---

$^{16}$The precise definition of this relation can be found in the Appendix 5.5.2.
5.2. Logic for games with intentions

together. But \( \mathcal{L}_{G,F} \) 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 \( \mathcal{L}_{G,F} \). By introducing it as a special propositional atom\(^{17}\), we can however capture in \( \mathcal{L}_{G,F} \) 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 \( w'(i) = s \).

\[
\mathbb{M}, w \models WR_i \quad \text{iff} \quad \forall s \in S_i \text{ such that } w(i) \neq s \text{ there is a } w' \sim_i w \text{ such that } w' \succeq_i w'[s/w'(i)] .
\]

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 \( \sigma \) is a Nash equilibrium if it satisfies the following:

\[
WR_1 \land K_1 \sigma(2) \land WR_2 \land K_2 \sigma(1)
\]

\(^{17}\)This notion could have been expressible in \( \mathcal{L}_{G,F} \), 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 $M, w \models WR_1$, we will be done if we can show that $[w]_i = \{w\}$. Now observe that $M, w \models K_1 \sigma(2)$ is just the same as saying that for all $w' \sim_1 w$, $M, w' \models \bigvee_{\sigma(2) = w(2)} \sigma$. 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 $L_{GF}$ 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. 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 $L_{GF}$ 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 $\Lambda_{L_{GF}}$ 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.

\[^{18}\text{As the reader may have come to realize, most of this high expressive power comes from the “hybrid” fragment, i.e. the constants that directly name strategy profiles. I show in Appendix 5.5.3 that nominals are indeed crucial to capture Nash equilibria.}\]
• All propositional tautologies.

• $S4$ for $\Box \leq$ and, for all $\sigma$ and $\sigma'$:
  \[(\text{Tot}) \quad (\sigma \land \Box i \leq \sigma') \lor (\sigma' \land \Box i \leq \sigma)\]

• For $\Diamond <, K$ and:
  \[(\text{Irr}) \quad \sigma \rightarrow \neg \Diamond < \sigma\]
  \[(\text{Trans}) \quad \Diamond < \Diamond < \sigma \rightarrow \Diamond < \sigma\]

• For $I_i$:
  \[(\text{K}) \quad I_i (\phi \land \psi) \leftrightarrow I_i \phi \land I_i \psi\]
  \[(\text{Ser}) \quad i_i T\]

• $S5$ for $E$.

• Interaction axioms.
  \[(\text{Exists}_\sigma) \quad <> \phi \rightarrow E\phi\]
  \[(\text{Inc}_{E, \neg \sigma}) \quad E(\sigma \land \phi) \rightarrow A(\sigma \rightarrow \phi)\]
  \[(\text{Inc}_\sigma) \quad E(\sigma)\]
  \[(\text{Inc}_1) \quad \text{As in Section 5.1.3.}\]
  \[(\text{K-I}) \quad I_i \phi \rightarrow K_i I_i \phi\]

• (Nec) for all modal connective, and the following additional inference rules. In both cases $\sigma \neq \sigma'$ and the former does not occur in $\phi$.

  – (Name) From $\sigma \rightarrow \phi$ infer $\phi$.
  – (Paste) From $(E(\sigma' \land <> \sigma) \land E(\sigma \land \phi)) \rightarrow \psi$ infer $E(\sigma' \land <> \phi) \rightarrow \psi$

Table 5.2: The axiom system for $\Lambda_L_{X,F}$. Here $<>_i$ is any of $\Box_i$, $\Diamond_i <$, $\Diamond_i \leq$ 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 $L_{X,F}$, while at the same time giving a more general perspective on transformations of epistemic game frames.
Chapter 5. Logics for practical reasoning with intentions

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 \(\phi\) means looking only at the sub-model where \(\phi\) holds. A public announcement formula, denoted \([\phi]!\psi\), thus says that after removing from the original models all the states where \(\phi\) does not hold, \(\psi\) is the case. As we shall see shortly, cleaning will indeed correspond to the announcement of a particular formula of \(L_{GF}\). But before showing this, let me look at the “public announcement” extension of \(L_{GF}\) in full generality.

5.3.1. Definition. [Dynamic extension of \(L_{GF}\)] \(DL_{GF}\), the dynamic extension of \(L_{GF}\), 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 \(\phi\), it is the case that \(\psi\)”. As I have just written, these announcements correspond to contractions of the underlying epistemic game model, that I denote \(M_{|\phi}\).

5.3.2. Definition. [Contracted epistemic game models] Given an epistemic game model \(M\) and a formula \(\phi \in DL_{GF}\), the contracted model \(M_{|\phi}\) is defined as follows\(^{20}\).

1. \(W_{|\phi} = ||\phi||\).
2. \(\sim_{i|\phi}\) is the restriction of \(\sim_i\) to \(W_{|\phi}\). Similarly for \(\geq_{i}\).
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 superset of \(A\), that is \(\uparrow A = \{B : A \subseteq B \subseteq W\}\). In item (3) the closure is with respect to \(W_{|\phi}\).
5.3. Transformations of games with intentions

4. $V_{\phi}$ is the restriction of $V$ to $W_{\phi}$.

The domain $W_{\phi}$ of a model restricted to $\phi$ is just what one would expect. It is the set of states where $\phi$ was the case before the announcement. The epistemic and preferences relations are modified accordingly.

The restriction of the intention function $\iota_i$ splits into two cases. On the one hand, if what is announced was compatible with the agent’s intention, that is if $\|\phi\| \cap \downarrow \iota_i(w) \neq \emptyset$, 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: $\iota_{\phi}(w) = \uparrow (\|\phi\| \cap \downarrow \iota_i(w))$. 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.

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 $M_{\phi}$ restricted after the announcement of $\phi$ is thus built out of the sub-model of $M$ where $\phi$ 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 5.7: The intention restriction when $\downarrow \iota_i(w) \cap \|\phi\| \neq \emptyset$.](image-url)
5.3.3. **Definition.** [Truth for public announcement formulas]

\[ M, w \models [\phi!] \psi \text{ iff } M, w \models \phi \text{ then } M_{\phi}, w \models \psi. \]

The condition “If \( M, w \models \phi \) 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 \( DLGF \) 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. 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 \( DLGF \). 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 \( \Lambda_{L_P} \) and \( \Lambda_{L_{GF}} \) 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 \( \lozenge \leq \lozenge \leq \phi \rightarrow \lozenge \leq \phi \) for transitivity of \( \geq \), together

\[^{21}\text{See van Benthem [2006a] for more on this phenomenon.}\]
5.3. Transformations of games with intentions

with some inference rules. The completeness arguments for these logics were also quite standard.

The axiomatization of the valid formulas of $DL_{GF}$ over the class of epistemic game frames proceeds differently. One does not need directly to provide formulas that correspond to properties of public announcements. 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 $L_{GF}$. If one can show that these formulas are valid, completeness of $\Lambda_{DL_{GF}}$ with respect to the class of epistemic game frames is just a corollary of completeness of $\Lambda_{L_{GF}}$ with respect to this very class of frames. By taking these formulas as axioms of $\Lambda_{L_{GF}}$, valid formulas in $DL_{GF}$ are then deductively reducible to valid formulas of $L_{GF}$, which we know can in turn be deduced in $\Lambda_{L_{GF}}$. 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 $\phi$, $i$ intends that $\psi$. Not surprisingly, these conditions match the two cases of the update rule for $\iota$. If the intentions of $i$ were already compatible with the announcements, that is if $i_\iota \phi$, 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 $\phi$ was not compatible with $\phi$, i.e. if $\neg i_\iota \phi$, then $i$ intends $\psi$ after the announcement if and only if $\psi$ is true everywhere in the restricted model, i.e. $[\phi!]A\psi$, which is exactly what the intention revision rule for $\iota$ prescribes.


color{black}

<table>
<thead>
<tr>
<th>Condition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $[\phi!]x \leftrightarrow \phi \rightarrow x$ for $x \in PROP \cup S$.</td>
<td></td>
</tr>
<tr>
<td>2. $[\phi!]-\psi \leftrightarrow \phi \rightarrow \neg[\phi!]\psi$.</td>
<td></td>
</tr>
<tr>
<td>3. $[\phi!]\psi \land \xi \leftrightarrow \phi \rightarrow ([\phi!]\psi \land [\phi!]\xi)$.</td>
<td></td>
</tr>
<tr>
<td>4. $[\phi!][.]\psi \leftrightarrow \phi \rightarrow [.]\phi \rightarrow [\phi!]\psi)$.</td>
<td></td>
</tr>
<tr>
<td>5. $[\phi!]I_i\psi \leftrightarrow \phi \rightarrow (i_\iota \phi \land I_i(\phi \rightarrow [\phi!]\psi) \lor (\neg i_\iota \phi \land [\phi!]A\psi))$</td>
<td></td>
</tr>
</tbody>
</table>

color{black}

Table 5.3: The axiom system for $\Lambda_{DL_{GF}}$. Here $[.]$ is either $A$, $K_i$, $\square_i^\leq$ or $\square_i^<$.  

\footnote{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 $\mathcal{L}_{GF}$ 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 $\mathcal{D}\mathcal{L}_{GF}$. 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 $\mathcal{D}\mathcal{L}_{GF}$, 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 $G$ and an intention profile $\iota$, the cleaned strategy set $cl(S_i)$ for an agent $i$ is defined as

$$cl(S_i) = \{s_i \mid \text{there is a } w' \in \downarrow \iota_i \text{ such that } w'(i) = s_i\}$$
5.3. Transformations of games with intentions

The altruistically cleaned version of $G$ from the point of view of the intention profile $\iota$, denoted $cl_\iota(G)$, is defined as follows.

- $cl(W) = \{w | \exists i \text{ such that } w(i) \in cl(S_i)\}$.
- $\sim_{i}^{cl}, \succeq_{i}^{cl}$ and $V^{cl}$ are restriction of $\sim_{i}, \succeq_{i}$ and $V^{cl}$ to $cl(W)$, respectively.
- For all $i$, $\iota_{i}^{cl} = \iota_{i}(cl(W) \cap \downarrow i_{i})$.

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 $G$ 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 $DL_{gf}$. It corresponds to the public announcement of a crucial concept: intention-rationality.

5.3.6. FACT. For any model $M$ 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 $w'' \in \downarrow i_{i}$ such that $w''(i) = w'(i)$, which is also the same as to say that there is an $i$ such that $M, w' \models \bigvee_{\sigma=\iota_{i}(\iota_{i}(\iota_{i}(\bigvee_{i \in I} IR_i)))}$. This last condition is equivalent to $M, w' \models \bigvee_{i \in I} IR_i$, 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 $\iota_{i}$:

$$\downarrow i_{i} \cap \bigvee_{i \in I} IR_i \neq \emptyset$$

Proof. Take any such $\iota_{i}$. We know that $\downarrow i_{i} \neq \emptyset$. So take any $w \in \downarrow i_{i}$. We have that $M, w \models i_{i}\sigma$ for $V(\sigma) = w$. But since we are working with state independent intentions, $i_{i}\sigma$ implies $\bigvee_{i \in I} IR_i$, 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 $\downarrow i_{i} \cap ||\bigvee_{i \in I} IR_i|| \neq \emptyset$. ■

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 intentions, 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 $\phi$ is said to be self-fulfilling at a pointed model $M, w$ if $M, w \models [\phi] A \phi$.

5.3.9. Fact. The announcement of $\bigvee_{i \in I} IR_i$ is self-fulfilling for any pointed model $M, w$ with state-independent intentions.

Proof. We have to show that for any pointed model with state-independent intentions, if $M, w \models \bigvee_{i \in I} IR_i$ then $M_{IR_i} \models A \bigvee_{i \in I} IR_i$. I will show something stronger, namely that for all $w \in ||IR_i||$, $M_{IR_i} \models A IR_i$.

Take any $w' \in W_{IR_i}$. We have to show that there is a $w''$ in $\downarrow_{IR_i}(w')$ 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 $\downarrow_{IR_i}(w')$ 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 $\downarrow_{IR_i}(w') = \downarrow_{IR_i}(w') \cap ||IR_i||$, and so that $w'' \in \downarrow_{IR_i}(w')$, 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).
5.3. Transformations of games with intentions

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 $\downarrow_{i}^{\mathbb{M}}(w') \cap [w']_i \neq \emptyset$. 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 $\downarrow_{i}^{\mathbb{M}_{\bigvee_{i \in I} IR_i^*}}(w') \cap [w']_i \subseteq \downarrow_{i}^{\mathbb{M}, w} IR_i^*$. This means that $w'' \in \downarrow_{i}^{\mathbb{M}_{\bigvee_{i \in I} IR_i^*}}(w') \cap [w']_i$, as required. $\blacksquare$

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:

$$\left\| \bigvee_{i \in I} IR_i^* \right\|_{\bigvee_{i \in I} IR_i^*} \subseteq \left\| \bigvee_{i \in I} IR_i \right\|_{\bigvee_{i \in I} IR_i}$$

Furthermore, there exist pointed models where this inclusion is strict.
Chapter 5. Logics for practical reasoning with intentions

**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 \( M, w_2 \models \neg IR_1^* \land \neg IR_2^* \). ■

**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 \( M, w \).

**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 \( \iota_i(w^\prime) = \iota_i(w^\prime) \). This was already ensured by the fact that \( w^\prime \sim_i w^\prime \). ■

**5.3.13. Fact.** The announcement of \( \bigvee_{i \in I} IR_i \) is not self-fulfilling for arbitrary pointed model \( M, w \).

**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
5.3. Transformations of games with intentions

<table>
<thead>
<tr>
<th>w</th>
<th>(i_1(w))</th>
<th>(i_2(w))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w_1)</td>
<td>(w_2, w_4)</td>
<td>(w_3, w_4)</td>
</tr>
<tr>
<td>(w_2)</td>
<td>(w_3)</td>
<td>(w_1)</td>
</tr>
<tr>
<td>(w_3)</td>
<td>(w_2)</td>
<td>(w_4)</td>
</tr>
<tr>
<td>(w_4)</td>
<td>(w_4)</td>
<td>(w_4)</td>
</tr>
</tbody>
</table>

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. \(M \models \bigwedge_{i \in I} [\phi!]_i \top\) for all models for game structure \(M\).

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 \(\bigwedge_{i \in I} [\phi!]_i \top\) is valid. In this proof the numbers refer to Table 5.3 on page 107.

We start with \([\phi!] \neg I_\perp\), which is the same as \([\phi!]_i \top\). By (2), this is equivalent to:

\[\phi \rightarrow \neg [\phi!]_i \perp\]

Now, by (5), the consequent expends into two parts \(\Phi = i_! \phi \land I_i (\phi \rightarrow [\phi!] \perp)\) and \(\Psi = \neg i_! \phi \land [\phi!] \top\), that I treat separately to keep the formulas readable.

\[\phi \rightarrow \neg (\phi \rightarrow (\Phi \lor \Psi))\]

Before looking at each disjunct, some redundancy can be eliminated by propositional reasoning, to get:

\[\phi \rightarrow \neg (\Phi \lor \Psi)\]

Now let us first look at \(\Phi = i_! \phi \land I_i (\phi \rightarrow [\phi!] \perp)\). By (1)—\(\perp\) can be treated as a propositional atom—we get:

\[i_! \phi \land I_i (\phi \rightarrow (\phi \rightarrow \perp))\]

This is equivalent in propositional logic to:

\[i_! \phi \land I_i (\neg \phi)\]

But the second conjunct is just the negation of the first, which means that \(\Phi\) is just equivalent to \(\perp\). We are thus left with:

\[\phi \rightarrow \neg (\perp \lor \Psi)\]
Which is just the same as:
\[ \phi \rightarrow \neg \Psi \]

Now, recall that \( B \) is the following:
\[ \neg i_i \phi \land [\phi!] A \bot \]

By (4), this expands to:
\[ \neg i_i \phi \land A(\phi \rightarrow [\phi!] \bot) \]

By (1) again, we thus get:
\[ \neg i_i \phi \land A(\phi \rightarrow (\phi \rightarrow \bot)) \]

This again reduces to:
\[ \neg i_i \phi \land A(\neg \phi) \]

Putting this back in the main formula, we get:
\[ \phi \rightarrow \neg (\neg i_i \phi \land A(\neg \phi)) \]

But then propositional reasoning gets us:
\[ (\phi \land A \neg \phi) \rightarrow i_i \phi \]

But the antecedent is just a contradiction of the axiom (Ref) for \( E \), and so we get:
\[ \bot \rightarrow i_i \phi \]

Which is a tautology. Since we took an arbitrary \( i \), we can conclude that \( \bigwedge_{i \in I} [\phi!] i_i \top \) 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 \( \phi \) at \( w \). The announcement of \( \phi \) 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 \( k \).
5.3. Transformations of games with intentions

Proof. Assume that there is no such \( k \).\(^{24} \) This means that there is no \( k \) such that \( M_k|_{\forall i \in I} w = M_{k+1}|_{\forall i \in I} w \). Since we work with finite models, this means that there is a finite \( n \)-step loop where \( M_k|_{\forall i \in I} w \neq M_{k+1}|_{\forall i \in I} w \neq \ldots \neq M_{k+n}|_{\forall i \in I} w \neq M_{k+n+1}|_{\forall i \in I} w \)

Now, observe that by Definition 5.3.2:

\[
W^k|_{\forall i \in I} w \supseteq W^{k+1}|_{\forall i \in I} w \supseteq \ldots \supseteq W^{k+n}|_{\forall i \in I} w \supseteq W^{k+n+1}|_{\forall i \in I} w
\]

But since \( M_k|_{\forall i \in I} w = M_{k+n+1}|_{\forall i \in I} w \), all these inclusion are in fact equalities.

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 \( w \in W^{k+\ell}|_{\forall i \in I} w \) such that \( \iota_{k+\ell}|_{\forall i \in I} w \neq \iota_{k+\ell+1}|_{\forall i \in I} w \). But this cannot be, as the following two cases show, and so there cannot be such a loop.

1. Assume that:

\[
\bigvee_{i \in I} V_{i \in I} w \neq \emptyset
\]

This means that:

\[
\iota_{k+\ell+1}|_{\forall i \in I} w = \uparrow \left( \bigvee_{i \in I} V_{i \in I} w \right) = \iota_{k+\ell+1}|_{\forall i \in I} w
\]

But observe that, while \( W^{k+\ell+1}|_{\forall i \in I} w = W^{k+\ell}|_{\forall i \in I} w \):

\[
\left\| \bigvee_{i \in I} V_{i \in I} w \right\|^{k+\ell+1} = W^{k+\ell+1}|_{\forall i \in I} w
\]

This means that:

\[
\iota_{k+\ell+1}|_{\forall i \in I} w = \uparrow \left( \bigvee_{i \in I} V_{i \in I} w \right) = \iota_{k+\ell+1}|_{\forall i \in I} w = \iota_{k+\ell+1}|_{\forall i \in I} w
\]

So (1) cannot hold while:

\[
\iota_{k+\ell+1}|_{\forall i \in I} w \neq \iota_{k+\ell+1}|_{\forall i \in I} w
\]

\(^{24}\)If we could show that this announcement is a monotone map, i.e. if it were the case that \( M|_{\forall i \in I} w \subseteq M'|_{\forall i \in I} w \), provided that \( M \subseteq M' \), then we would be done. See Apt [2007]. Unfortunately this announcement is not monotonic. We indeed have that \( W|_{\forall i \in I} w \subseteq W'|_{\forall i \in I} w \), if \( M \subseteq M' \). The non-monotonicity lies in the update rule for the intention set. It is not very complicated to devise an example where \( M \subseteq M' \) but in which there is a \( w \in W \) and an \( i \in I \) such that \( \iota_{i|_{\forall i \in I} w} \neq \iota_{i|_{\forall i \in I} w} \). For this reason I show the existence of the fixed point directly.
2. Assume then that:
\[ \downarrow_{i|V_{i \in I}R_i}(w) \cap \bigvee_{i \in I} IR_i \bigg|^{|k+\ell} \ = \ \emptyset \]  

(2)

In this case \( \downarrow_{i|V_{i \in I}R_i}(w) \) just becomes \( \{W_{i|V_{i \in I}R_i}^{k+\ell+1}\} \). But recall that by definition, \( W_{i|V_{i \in I}R_i}^{k+\ell+1} \) is just \( \bigvee_{i \in I} IR_i \bigg|^{|k+\ell} \). But since we know that \( W_{i|V_{i \in I}R_i}^{k+\ell+1} = W_{i|V_{i \in I}R_i}^{k+\ell} \), this means that \( \bigvee_{i \in I} IR_i \bigg|^{|k+\ell} = W_{i|V_{i \in I}R_i}^{k+\ell} \). But that would mean:
\[ \downarrow_{i|V_{i \in I}R_i}(w) \cap W_{i|V_{i \in I}R_i}^{k+\ell} = \emptyset \]

Which is just to say that
\[ \downarrow_{i|V_{i \in I}R_i}(w) = \emptyset \]

Which is impossible by Fact 5.3.14.

\[ \boxed{} \]

5.3.17. COROLLARY. If the announcement of intention-rationality stabilizes at \( k \) for a given pointed model \( M, w \), then \( M|_{\bigvee_{i \in I} IR_i}, w |\bigvee_{i \in I} IR_i = \bigvee_{i \in I} IR_i. \)

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, \( M|_{\bigvee_{i \in I} IR_i}, w |\bigvee_{i \in I} IR_i \).

**Proof.** Assume not, then \( M|_{\bigvee_{i \in I} IR_i}, w |\bigvee_{i \in I} IR_i \). But then \( w \notin M^{k+1}_{i|V_{i \in I}R_i} \), against the assumption that the announcement of \( \bigvee_{i \in I} IR_i \) stabilizes at \( k \). \[ \boxed{} \]

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 \( M, w \), the announcement of \( \bigvee_{i \in I} IR_i \) stabilizes after one announcement.

**Proof.** By definition, \( W_{i|V_{i \in I}R_i} = \bigvee_{i \in I} IR_i \). But we also know from Fact 5.3.12 that for all \( w \) in \( W_{i|V_{i \in I}R_i} \), \( M|_{\bigvee_{i \in I} IR_i, w |\bigvee_{i \in I} IR_i} \). This means that \( M^2_{i|V_{i \in I}R_i} = M|_{\bigvee_{i \in I} IR_i}. \) \[ \boxed{} \]

\(^{25}\)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 $M$ where the announcement of intention-rationality stabilizes at $k$ such that:

$$M|_{V_{i \in I} IR_i^*} \not\subseteq M|_{V_{i \in I} IR_i}$$

Proof. Take a model $M$ 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

\[
\begin{array}{c|c|c}
  w & \downarrow_1(w) & \downarrow_2(w) \\
  \hline
  w_1 & w_2 & w_1 \\
  w_2 & w_1 & w_4 \\
  w_3 & w_3 & w_3 \\
  w_4 & w_4 & w_4 \\
\end{array}
\]

Table 5.5: The intentions of the agents in counterexample for Fact 5.3.20.

that in all states, $M, w \models \bigvee_{i \in I} IR_i^*$. This means that for all states, $M|_{V_{i \in I} IR_i}, w = M, w$, i.e. this announcement does not remove any states, and so that $M$ is its own stabilization point. But observe, on the other hand, that at $M, w_2 \not\models \bigvee_{i \in I} IR_i^*$. But since $\downarrow_1(w_1) = \{w_2\}$, we get $\downarrow_{1|_{V_{i \in I} IR_i^*}}(w_1) = \{w_1, w_3, w_4\}$ after the announcement of knowledge-consistency at $w_1$. But then it is clear that $\downarrow_{1|_{V_{i \in I} IR_i^*}}(w_1) \not\subseteq \downarrow_{1|_{V_{i \in I} IR_i}}(w_1)$, and since in this case the announcement of $\bigvee_{i \in I} IR_i$ "stabilizes" at $k = 0$, we get that $M|_{V_{i \in I} IR_i^*} \not\subseteq M|_{V_{i \in I} IR_i}$.

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 $M, w$, if $M, w \models \bigvee_{i \in I} IR_i^*$ then $M, w \models \bigvee_{i \in I} IR_i | \bigvee_{i \in I} IR_i^*$.

Proof. Assume that $M, w \models \bigvee_{i \in I} IR_i^*$, i.e. that there is an $i$ and a $w' \sim_i w$ such that $w' \in \downarrow_i(w)$. Because $IR_i^*$ is introspective, this means that $M, w' \models \bigvee_{i \in I} IR_i^*$. But then $M, w' \models \bigvee_{i \in I} IR_i$, which means that both $w'$ and $w$ are in $W|_{V_{i \in I} IR_i}$, and also that $w' \in \downarrow_{i|_{V_{i \in I} IR_i}}(w)$. But then $M, w \models \bigvee_{i \in I} IR_i | \bigvee_{i \in I} IR_i^*$.
5.3.22. **Corollary.** Suppose that for a pointed model $\mathcal{M}, w$ the announcement of intention-rationality stabilizes at $k$, then

$$\left\| \bigvee_{i \in I} IR_i^* \right\|_{\| \cdot \|_{IR_i}} \subseteq \left\| \bigvee_{i \in I} IR_i \right\|^k_{\| \cdot \|_{IR_i}}$$

Using $D\mathcal{L}_{GF}$, 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 $\mathcal{G}, w$, 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: $\mathcal{M}, w \models \bigvee_{\sigma \in S} \bigwedge_{i \in I} i, \sigma$. With this in hand, one can directly show general conditions for overlapping.

5.3.24. **Fact.** [Intention overlap] For any pointed model $\mathcal{M}, w$, the following are equivalent:

(i) $\mathcal{M}, w \models [\phi!] \bigvee_{\sigma \in \| \phi \|} \bigwedge_{i \in I} i, \sigma$

(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:

$$\left[\phi!\right] \bigvee_{\sigma \in \|\phi\|} \bigwedge_{i \in I} i_i \sigma$$

This formula is propositionally equivalent to:

$$\left[\phi!\right] \neg \bigwedge_{\sigma \in \|\phi\|} \neg \bigwedge_{i \in I} i_i \sigma$$

Then, by (2) and (3), we get:

$$\phi \rightarrow \neg \bigwedge_{\sigma \in \|\phi\|} \left[\phi!\right] \neg \bigwedge_{i \in I} i_i \sigma$$

By (2) again, we obtain:

$$\phi \rightarrow \neg \bigwedge_{\sigma \in \|\phi\|} \left(\phi \rightarrow \neg \bigwedge_{i \in I} \left[\phi!\right] \neg I_i \neg \sigma\right)$$

Now we can replace $i_i$ by its dual:

$$\phi \rightarrow \neg \bigwedge_{\sigma \in \|\phi\|} \left(\phi \rightarrow \neg \bigwedge_{i \in I} \left[\phi!\right] \neg I_i \neg \sigma\right)$$

We then reapply (2):

$$\phi \rightarrow \neg \bigwedge_{\sigma \in \|\phi\|} \left(\phi \rightarrow \neg \left[\phi!\right] \neg I_i \neg \sigma\right)$$

This, after some transformation from propositional logic, reduces to:

$$\phi \rightarrow \bigvee_{\sigma \in \|\phi\|} \bigwedge_{i \in I} \left(\phi \rightarrow \neg \left[\phi!\right] \neg I_i \neg \sigma\right)$$

Now I will look at $\left[\phi!\right] I_i \neg \sigma$ separately. By (5) it reduces to:

$$\phi \rightarrow \left((i_i \phi \land I_i (\phi \rightarrow \left[\phi!\right] \neg \sigma)) \lor (I_i \neg \phi \land \left[\phi!\right] A \neg \sigma)\right)$$

By (2) and (4), this is the same as:

$$\phi \rightarrow \left((i_i \phi \land I_i (\phi \rightarrow \neg \sigma)) \lor (I_i \neg \phi \land A (\phi \rightarrow \neg \sigma))\right)$$
Now, reinserting this formula in the main one and pushing the negation inside, we get:

$$\phi \rightarrow \bigvee_{\sigma \in ||\phi||} \bigwedge_{i \in I} (\phi \rightarrow (i \phi \rightarrow i_i(\phi \land \sigma)) \land (I_i \neg \phi \rightarrow E(\phi \land \sigma))))$$  \hspace{1cm} \text{(Post)}

I am now ready for the correspondence argument, which boils down to showing that (Post) is true at a pointed model $M, w$ iff (ii) holds. We have that $M, w, |\models \text{Post}$ iff there is a $\sigma$ and a $w' \in ||\phi||$ with $V(\sigma) = \{w'\}$ such that for all $i \in I$ the two conjuncts hold, provided that $M, w |\models \phi$. Observe first that the two conjuncts divide the agents into two groups. On the one hand are those such that $\downarrow i_i(w) \cap ||\phi|| \neq \emptyset$, i.e. those whose intentions are compatible with the announcement of $\phi$. 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 $I_i \neg \phi \rightarrow E(\phi \land \sigma)$. This simply restates what we already knew, namely that $\phi$ holds at $w'$, which means that the truth of this formula in fact only depends on the first conjunct, $i_i \phi \rightarrow i_i(\phi \land \sigma)$. This bluntly says that $w'$ was already compatible with the intentions of all the agents in the first group before the announcement of $\phi$, 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 $L_{GF}$ whose announcement would ensure intention overlap.

Fact 5.3.24 states that overlap occurs after an announcement whenever it occurred before, i.e. 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
5.3. Transformations of games with intentions

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 \( \phi \) enables the announcement of \( \psi \) for a given model \( M \) whenever the following holds:

- \( M|_{\psi} = M \)
- \( M|_{\phi \psi} \subset M|_{\phi} \subset M \)

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 \( \phi \) enables the announcement of \( \psi \) when information in an epistemic game frame is not affected by the announcement of \( \psi \) alone, but it is affected by this announcement after the announcement of \( \phi \) 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 \( M \), the announcement of \( WR_i \) enables the announcement of \( IR^*_j \) iff for all \( w, M, w \models IR_j \) but there are some \( w \) such that

- \( M, w \models WR_i \).
- \( \downarrow t_j(w) \cap ||WR_i|| \neq \emptyset \).
- For all \( w' \in \downarrow t_j(w) \cap [w]_j, M, w' \not\models WR_i \).
Page 122

Chapter 5. Logics for practical reasoning with intentions

Proof. First the left-to-right direction. Assume that in \(M\) the announcement of \(WR_i\) enables the announcement of \(IR_j\). That means first that \(M\models IR_j = M\), which means that for all \(w \in W, M, w \models IR_j\), i.e. \(\downarrow \iota_j(w) \cap [w]_j \neq \emptyset\). On the other hand, that \(M\models IR_j \subset \mathbb{M}\models IR_j\), means that for some of these \(w\), we have \(M, w \models WR_i\) and \(\downarrow \iota_j(w) \cap [w]_j = \emptyset\). 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 \(\iota_j|WR_i(w) = \{\downarrow \iota_j(w) \cap \|WR_i\|\}\). This means that there are some \(w' \subseteq \downarrow \iota_j(w) \cap [w]_j\), while \(w' \neq \downarrow \iota_j|WR_i(w) \cap [w]_j\|WR_i\). Now, unpacking the definitions of the update rule for the intention set and the epistemic relation, we get, for all \(w\):

\[
\text{if } w \in (\iota_i(w) \cap [w]_i \cap \|WR_j\|) \text{ then } w \in \downarrow \iota_i|WR_i(w) \cap [w]_i|WR_i,
\]

Putting this in contrapositive, we get that for all \(w' \not\subseteq \iota_j|WR_i(w) \cap [w]_j|WR_i\), while being in \(\iota_j(w) \cap [w]_j\), \(w' \not\subseteq \|WR_i\|\)

For the right-to-left direction, observe first that for all \(w, M, w \models IR_j\) is the same as to say that \(M\models IR_j = M\). Now, take one \(w\) as specified. Since \(M, w \models WR_i\) and \(M, w' \not\models WR_i\) for all \(w' \subseteq \iota_j(w) \cap [w]_j\), we know that \(M\models WR_i \subset M\) and that the former is not empty. Now, because \(\downarrow \iota_j(w) \cap \|WR_i\| \neq \emptyset\) we also know that \(\iota_j|WR_i(w) = \{\downarrow \iota_j(w) \cap \|WR_i\|\}\), which means that \(\downarrow \iota_j|WR_i(w) \subseteq \iota_j(w)\). Moreover \([w]_j|WR_i \subseteq [w]_j\) by definition. This means that \((\|WR_i\| \\downarrow \iota_j(w) \cap [w]_j) \subseteq \|\|WR_i\|\), which in the present context can only result in \(\downarrow \iota_j|WR_i(w) \cap [w]_j|WR_i\). This means that \(w \not\subseteq \|\|WR_i\|\), and so that \(M\models |WR_i|^{IR_j} \subset M\models |WR_i|\).

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. Transformations of games with intentions

5.3.27. **Fact.** [No self-enabling] If $M, w \models WR_j$ and $M, w' \models IR_i^*$ 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 $M_{\models |WR_j| |IR_i^*|} = M_{\models |WR_i|}$, provided that $M, w \models WR_i$ and $M, w' \models IR_i^*$ for all $w' \in W$. This, in turn, follows from the Lemma 5.3.28 (below).

5.3.28. **Lemma.** For any pointed model $M, w$, if $M, w \models IR_i^*$ and $M, w \models WR_i$ then $M, w \models [WR_i!] IR_i^*$.

**Proof.** I will show the contrapositive. The proof rests crucially on the fact that $WR_i$ is introspective. That is, for any $M, w \models WR_i$ we also have $M, w' \models WR_i$ for all $w' \sim_i w$.

Assume that $M, w \not\models [WR_i!] IR_i^*$ while $M, w \models WR_i$. That means that $M_{\models |WR_i|} |IR_i^*| \not\models IR_i^*$, i.e. that $\downarrow_{i|WR_i}(w) \cap [w]_{i|WR_i} = \emptyset$. Now, as usual, $\downarrow_{i|WR_i}(w)$ can come from two sources.

1. It can be that $\downarrow_{i|WR_i}(w) = \{W_{|WR_i|}\}$ because $\downarrow_{i}(w) \cap ||WR_i|| = \emptyset$. But because $M, w \models WR_i$ and $WR_i$ is introspective, we know that $[w]_i \subseteq ||WR_i||$, which means that $\downarrow_{i}(w) \cap [w]_i = \emptyset$, i.e. $M, w, \not\models IR_i^*$.

2. It thus remains to check what happens when $\downarrow_{i}(w) \cap ||WR_i|| \neq \emptyset$. I will show that in that case $\downarrow_{i}(w) \cap [w]_i$ and $\downarrow_{i|WR_i}(w) \cap [w]_{i|WR_i}$ 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 $w' \in \downarrow_{i}(w) \cap [w]_i$. 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 $\downarrow_{i}(w) \cap ||WR_i|| \neq \emptyset$, we also know that $\downarrow_{i|WR_i}(w) = \downarrow_{i}(w) \cap ||WR_i||$, which means that $w'$ is in $\downarrow_{i|WR_i}(w)$ as well.

So $\downarrow_{i}(w) \cap [w]_i = \downarrow_{i|WR_i}(w) \cap [w]_{i|WR_i}$. The required conclusion is then a direct consequence of our assumption that $\downarrow_{i|WR_i}(w) \cap [w]_{i|WR_i} = \emptyset$.

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 Beuthem [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 $DL_{\forall \forall}$. 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\textsuperscript{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 $L_P$

5.5.1. Definition. [Modal equivalence] Two pointed preference models $M, w$ and $M', v$ are modally equivalent, noted $M, w \sim\approx M', v$, iff for all formula $\phi$ of $L_P$, $M, w \models \phi$ iff $M', v \models \phi$.

5.5.2. Definition. [Bisimulation] Two pointed preference models $M, w$ and $M', v$ are bisimilar, noted $M, w \leftrightarrow M', v$, whenever there is a relation $E \subseteq M \times M'$ 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 $M, w$ and $M', v$. If $E$ is a bisimulation and $M, w \sim\approx M', v$, then we say that $w$ and $v$ are bisimilar, which is noted $w \sim\leftrightarrow v$.

\textsuperscript{27}Key references here are Cohen and Levesque [1990], Georgeff et al. [1998] and Wooldridge [2000].
Chapter 5. Logics for practical reasoning with intentions

As usual in modal logic, one can show that any two bisimilar pointed preference models are modally equivalent. In other words, truth in $\mathcal{L}_P$ is invariant under bisimulation.

### 5.5.2 More on lifted relations in $\mathcal{L}_P$

The following lifted preference relations can be defined in $\mathcal{L}_P$:

**5.5.3. Definition.** [Binary preference statements]

1. $\psi \geq_{\exists\exists} \phi \iff E(\phi \land \diamond \leq \psi)$
2. $\phi \leq_{\forall\exists} \psi \iff A(\phi \rightarrow \diamond \leq \psi)$
3. $\psi >_{\exists\exists} \phi \iff E(\phi \land \diamond < \psi)$
4. $\phi <_{\forall\exists} \psi \iff A(\phi \rightarrow \diamond < \psi)$

The formulas $\psi \geq_{\exists\exists} \phi$ and $\psi >_{\exists\exists} \phi$ may be read as “there is a $\psi$-state that is at least as good as a $\phi$-state” and “there is a $\psi$-state that is strictly better than a $\phi$-state”, respectively. The other comparative statements, $\phi \leq_{\forall\exists} \psi$ and $\phi <_{\forall\exists} \psi$, can be read as “for all $\phi$-states there is an at least as good $\psi$-state” and as “for all $\phi$-state there is a strictly preferred $\psi$-state”, respectively.

We can define further binary preference statements as duals of the above modalities.

**5.5.4. Definition.** [Duals]

5. $\phi >_{\forall\forall} \psi \iff \neg(\psi \geq_{\exists\exists} \phi) \iff A(\phi \rightarrow \Box \leq \neg \psi)$
6. $\phi >_{\exists\forall} \psi \iff \neg(\psi \leq_{\forall\exists} \phi) \iff E(\phi \land \Box \leq \neg \psi)$
7. $\phi \geq_{\forall\forall} \psi \iff \neg(\psi >_{\exists\exists} \phi) \iff A(\phi \rightarrow \Box < \neg \psi)$
8. $\phi \geq_{\exists\forall} \psi \iff \neg(\psi <_{\forall\exists} \phi) \iff E(\phi \land \Box < \neg \psi)$

The first formula tells us that “everywhere in the model, if $\phi$ is true at a state, then there is no $\psi$-state at least as good as it”. Observe that if the underlying preference relation is total, this boils down to saying that all $\phi$-states, if any, are strictly preferred to all $\psi$-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 $\phi$-state strictly preferred to all the $\psi$-states, if any. These intended meaning are, however, not definable in $\mathcal{L}_P$ 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 $\mathcal{L}_P$ 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 \( \mathbb{M} \) if whenever \( \geq \) 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 \( \geq \) for satisfied formulas and with \( \leq \) for any formulas. Transitivity does lift with \( \leq \) for satisfied formulas, but not with \( \geq \).

**Proof.** The proof is trivial for reflexivity, in both cases, for totality with \( \geq \) and for transitivity with \( \leq \), both for satisfied formulas. Totality for \( \leq \) has been proved in Section 5.1.2.

For failure of transitivity lifting with \( \geq \), take a preference model with four states where \( w_1 \geq w_2 \geq w_3 \geq w_2 \). Make \( \phi \) only true at \( w_1 \) and \( w_4 \), \( \psi \) only at \( w_3 \) and \( \xi \) only at \( w_3 \). We clearly get \( \psi \geq \xi \), \( \phi \geq \xi \) but not \( \psi \geq \xi \). ■

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 \( \mathcal{L}_{GF} \)**

In the following definition, the clauses 2 and 3 are intended to apply to both \( \geq \) and \( \sim \). 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 \( \mathbb{M}, w \leftrightarrow \mathbb{M}', v \), whenever there is a relation \( E \subseteq \mathbb{M} \times \mathbb{M}' \) such that:

1. For all \( x \in \text{PROP} \cup S \), \( w \in V(x) \) iff \( v \in V(x) \),
2. (Forth -R) if \( w' \mathrel{R} w \) then there is a \( v' \in W' \) such that \( v' \mathrel{R} v \) and \( w' \mathrel{E} v' \),
3. (Back -R) if \( v' \mathrel{R} v \) then there is a \( w' \in W \) such that \( w' \mathrel{R} w \) and \( v' \mathrel{E} w' \),
4. [ten Cate, 2005, p.47] For all \( \sigma \in S \), if \( V(\sigma) = \{ w \} \) and \( V'(\sigma) = \{ v \} \) then \( w \mathrel{E} v \).
5. [Hansen, 2003, p.18] (Forth \( -\iota \)) if \( X \in \iota(w) \) then there is an \( X' \subseteq W' \) such that \( X' \in \iota(v) \) and for all \( v' \in X' \) there is a \( w' \in X \) such that \( w' \mathrel{E} v' \).
6. [Hansen, 2003, p.18] (Back \( -\iota \)) if \( X' \in \iota(v) \) then there is an \( X \subseteq W \) such that \( X \in \iota(w) \) and for all \( w' \in X' \) there is a \( v' \in X \) such that \( v' \mathrel{E} w' \).

Truth in \( \mathcal{L}_{GF} \) is indeed invariant under this notion of bisimulation.

5.5.9. **Fact.** [Nash Equilibrium without nominals] Let \( \mathcal{L}_{GF}^- \) be \( \mathcal{L}_{GF} \) minus the nominals. Nash equilibrium is not definable in \( \mathcal{L}_{GF}^- \).
Table 5.6: Two games with bisimilar models but different Nash equilibria. The payoffs are identical for both players.

<table>
<thead>
<tr>
<th>G1</th>
<th>t1</th>
<th>t2</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>s2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>G2</th>
<th>t0</th>
<th>t1</th>
<th>t2</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>s2</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**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 $i \in \{1, 2\}$, $t_i(w) = \{W\}$, $\{w' : w \sim_i w'\} = \{w\}$ and for all $p \in \text{prop}$, $V(p) = \emptyset$, and similarly in $M'$.

It is easy to check that $L_{GF}^-$ is invariant under bisimulation as just defined, without clauses related to nominals. $M$ and $M'$ 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.

![Fig. 5.10](image)

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 $L_{GF}$. 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 $L_{GF}$. 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 expressible power provided by nominals. In such models all properties definable by a pure formula, i.e. a formula with only nominals as atoms, are canonical (see ten Cate [2005, p.69]). During the construction of
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 $\Lambda_{\mathcal{L}_GF}$, by proving existence lemmas for $K_i$, $\lozenge <$, $\lozenge \leq$ 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 $\mathcal{L}_{GF}$, 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 $\Gamma$ be a maximally consistent set (MCS) of $\Lambda_{\mathcal{L}_{GF}}$. We say that $\Gamma$ is *named* by $\sigma$ if $\sigma \in \Gamma$. If $\sigma$ names some MSC(s) $\Gamma$ we denote it (them) $\Gamma_{\sigma}$. $\Gamma$ is *pasted* whenever $E(\sigma \land <> \phi) \in \Gamma$ implies that $E(\sigma \land <> \sigma') \land E(\sigma' \land \phi)$ is also in $\Gamma$.

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 $\mathcal{L}_{GF}'$ be $\mathcal{L}_{GF} \cup S'$. Then every $\Lambda_{\mathcal{L}_{GF}}$ consistent set of formulas can be extended to a named and pasted $\Lambda_{\mathcal{L}_{GF}'}$-MCS.

<table>
<thead>
<tr>
<th>Profile in $G_1$</th>
<th>State in $M$</th>
<th>bisimilar to in $M'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(s_1, t_1)$</td>
<td>$v_1$</td>
<td>$w_1, w_2, w_3, w_5, w_6$</td>
</tr>
<tr>
<td>$(s_1, t_2)$</td>
<td>$v_2$</td>
<td>$w_2$</td>
</tr>
<tr>
<td>$(s_2, t_1)$</td>
<td>$v_3$</td>
<td>$w_1, w_3, w_5$</td>
</tr>
<tr>
<td>$(s_2, t_2)$</td>
<td>$v_4$</td>
<td>$w_4$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Profile in $G_2$</th>
<th>State in $M'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(s_1, t_1)$</td>
<td>$w_1$</td>
</tr>
<tr>
<td>$(s_1, t_2)$</td>
<td>$w_2$</td>
</tr>
<tr>
<td>$(s_2, t_1)$</td>
<td>$w_3$</td>
</tr>
<tr>
<td>$(s_2, t_2)$</td>
<td>$w_4$</td>
</tr>
<tr>
<td>$(s_1, t_0)$</td>
<td>$w_5$</td>
</tr>
<tr>
<td>$(s_2, t_0)$</td>
<td>$w_6$</td>
</tr>
</tbody>
</table>

Table 5.7: The bisimulation for the models in Figure 5.10.
Chapter 5. Logics for practical reasoning with intentions

Proof.

Naming Enumerate $S'$, and let $\sigma$ be the first new nominal in that enumeration. For a given consistent set $\Gamma^*$, fix $\Gamma_\sigma = \Gamma \cup \{\sigma\}$. By (Name) $\Gamma_\sigma$ is consistent.

Pasting Enumerate the formulas of $\mathcal{L}_{\mathcal{G}F}'$ and take $\Gamma_0 = \Gamma_\sigma$. Assume $\Gamma_n$ is defined, and let $\phi_{n+1}$ be the $n^{th} + 1$ formula in the enumeration. Define $\Gamma_{n+1}$ as $\Gamma_n$ if $\Gamma_n \cup \{\phi_{n+1}\}$ is inconsistent. Otherwise form $\Gamma_{n+1}$ by adding $\phi_{n+1}$ to $\Gamma_n$ if $\phi_{n+1}$ is not of the form $E(\sigma' \wedge \psi)$. If $\phi_{n+1}$ is of form $E(\sigma' \wedge < > \sigma'') \wedge E(\sigma'' \wedge \phi)$, then we paste with the first new nominal $\sigma''$ 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), $\Gamma_{n+1}$ is also consistent. Set, finally, $\Gamma = \bigcup_{n \leq \omega} \Gamma_n$. This is clearly a named and pasted MCS.

$\blacksquare$

5.5.12. Definition. [Yielded MCS] The sets yielded by a $\Lambda_{\mathcal{L}_{\mathcal{G}F}'}$-MCS $\Gamma$ are the sets $\Delta_\sigma$ 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 $\Delta_\sigma$ and $\Delta_{\sigma'}$ be any yielded sets of a $\Lambda_{\mathcal{L}_{\mathcal{G}F}'}$-MCS $\Gamma$, for arbitrary nominals $\sigma$ and $\sigma'$ in $\mathcal{L}_{\mathcal{G}F}'$.

1. Both $\Delta_\sigma$ and $\Delta_{\sigma'}$ are named $\Lambda_{\mathcal{L}_{\mathcal{G}F}'}$-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 $\sigma''$ names $\Gamma$ then $\Gamma$ is itself the yielded set $\Delta_{\sigma''}$.

Proof.

1. By (Exists$\sigma$), $E\sigma \in \Gamma$, and thus $\Delta_\sigma$ is named. Assume now it is not consistent. That means that there are $\xi_1 \wedge \ldots \wedge \xi_n$ such that one can prove $\neg(\xi_1 \wedge \ldots \wedge \xi_n)$ in $\Lambda_{\mathcal{L}_{\mathcal{G}F}}$. But that means that $\neg A(\xi_1 \wedge \ldots \wedge \xi_n) \in \Gamma$, by (Nec). This, in turns, means that $\neg E(\xi_1 \wedge \ldots \wedge \xi_n) \in \Gamma$. But that can’t be. Recall that $(\xi_1 \wedge \ldots \wedge \xi_n) \in \Delta_\Gamma$ iff $E(\sigma \wedge \xi_1 \wedge \ldots \wedge \xi_n)$ is also in $\Gamma$. But then by (K) for $E$, we get that $E(\xi_1 \wedge \ldots \wedge \xi_n) \in \Gamma$. For maximality, observe that a formula $\phi$ and its negation are not in $\Delta_\sigma$ iff neither $E(\sigma \wedge \phi)$ nor $E(\sigma \wedge \neg \phi)$ are in $\Gamma$. 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 $\Gamma$. 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 $\Gamma$. 
2. Assume \( \sigma' \in \Delta_{\sigma} \). That means that \( E(\sigma \land \sigma') \in \Gamma \). By (Inc\(_E\)\(_{-\sigma}\)) we get that both \( A(\sigma \rightarrow \sigma') \) and \( A(\sigma' \rightarrow \sigma) \) are in \( \Gamma \), and so by \( K \) for \( E \), we get \( A(\sigma \rightarrow \sigma') \in \Gamma \). Assume now that \( \phi \in \Delta_{\sigma} \). This means that \( E(\sigma \land \phi) \in \Gamma \). But standard \( K \) reasoning we get that \( E(\sigma' \land \phi) \in \Gamma \), which means that \( \phi \) 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 \( E(\sigma' \land \phi) \in \Delta_{\sigma} \). This means that \( E(\sigma \land E(\sigma' \land \phi)) \in \Gamma \). But then this implies, by \( K \) for \( E \) that \( EE(\sigma' \land \phi) \in \Gamma \), which in turns, because of (Trans) for \( E \), implies \( E(\sigma' \land \phi) \in \Gamma \). For the converse, assume that \( E(\sigma' \land \phi) \in \Gamma \). By (Sym) for \( E \), we get that \( AE(\sigma' \land \phi) \in \Gamma \). But we also know by (Exists\(_{\sigma}\)) that \( E\sigma \in \Gamma \), from which we get by standard \( K \) reasoning that \( E(\sigma \land E(\sigma' \land \phi)) \in \Gamma \). This means that \( E(\sigma' \land \phi) \in \Delta_{\sigma} \).

4. Assume that \( \sigma \in \Gamma \). For the left to right, assume that \( \phi \in \Gamma \). This means that \( \sigma \land \phi \in \Gamma \), which implies by (Ref) that \( E(\sigma \land \phi) \) and so that \( \phi \in \Delta_{\sigma} \). Now assume that \( \phi \in \Delta_{\sigma} \). This means that \( E(\sigma \land \phi) \in \Gamma \), which in turn implies that \( A(\sigma \rightarrow \phi) \) by (Inc\(_E\)\(_{-\sigma}\)). But then by (Ref) again we get that \( \sigma \rightarrow \phi \in \Gamma \), and \( \phi \) itself because \( \sigma \in \Gamma \).

\[ \blacksquare \]

This last lemma prepares the ground for the hybrid fragment for \( \Lambda_{\mathcal{L}} \). 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 \( f : W \rightarrow \mathcal{P}(\mathcal{P}(W)) \) is:

- **closed under superset**s 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 superset and under binary intersection.

5.5.15. DEFINITION. [Neighbourhood tools] [Pacuit, 2007, p.8-9]
Let \( W^\Gamma \) be the set of all named sets yielded by \( \Gamma \).

- The **proof set** \( |\phi| \) of a formula \( \phi \) of \( \mathcal{L}_{\mathcal{G}F}' \) is defined as \( \{ \Delta_{\sigma} \in W^\Gamma : \phi \in \Delta_{\sigma} \} \).
- A neighbourhood function \( \iota \) is **canonical** for \( \Lambda_{\mathcal{L}} \) if for all \( \phi \), \( |\phi| \in \iota(\Delta_{\sigma}) \) iff \( I_\iota \phi \in \Delta_{\sigma} \).
- A neighbourhood function \( \iota_{\text{min}} : W^\Gamma \rightarrow \mathcal{P}(\mathcal{P}(W)) \) is **minimal** if \( \iota_{\text{min}}(\Delta_{\sigma}) = \{ |\phi| : \phi \in \Delta_{\sigma} \} \).
• The supplementation $\uparrow t_{\min}$ of $t_{\min}$ is the smallest function that contains $t_{\min}(\Delta_\sigma)$ and that is closed under supsersets.

5.5.16. FACT. [Properties of $\uparrow t_{\min}$] [Pacuit, 2007] $\uparrow t_{\min}$ is well-defined, canonical for $\Lambda_{\mathcal{L}_\sigma^F}$ and a filter.

5.5.17. DEFINITION. [Epistemic model for completeness] Let $\Gamma$ be any named and pasted $\Lambda_{\mathcal{L}_\sigma^F}$-MCS. The named game model $M^\Gamma$ yielded by $\Gamma$ is a tuple $\langle W^\Gamma, I, \sim^\Gamma_i, \succeq^\Gamma_i, \succ^\Gamma_i, \iota^\Gamma_i, V^\Gamma \rangle$ such that:

- $W^\Gamma$ is the set of sets yielded by $\Gamma$.
- $I$, defined as $\{ i : \text{there is a } < >_i \phi \text{ in } \mathcal{L}_\sigma^F \}$, is the set of agents.
- $\Delta_\sigma \sim^\Gamma_i \Delta_\sigma'$ iff for all $\phi \in \Delta_\sigma'$, $\Diamond^\Gamma_i \phi \in \Delta_\sigma$, and similarly for $\succeq^\Gamma_i$ and $\succ^\Gamma_i$.
- $\iota^\Gamma_i(\Delta_\sigma) = \uparrow \iota^\Gamma_i,\text{min}(\Delta_\sigma)$.
- 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 $E\phi$, $K_i$, $\Diamond^\leq_i$ and $\Diamond^<_i$). If $\Diamond^\leq_i \phi \in \Delta_\sigma$ then there is a $\Delta_\sigma' \in W$ such that $\phi \in \Delta_\sigma'$ and $\Delta_\sigma \sim^\Gamma_i \Delta_\sigma'$. Similarly for $\Diamond^<_i$ and $E\phi$. Furthermore, if $\phi \in \Delta_\sigma$ then for all $\Delta'_\sigma$, $E\phi \in \Delta'_\sigma$.

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 $|\phi| \in \iota^\Gamma_i(\Delta_\sigma)$.

Proof. Trivially follows from the definition of $\iota^\Gamma_i$. ■

5.5.20. LEMMA (Truth Lemma). For all $\phi \in \Gamma$, $M^\Gamma, \Delta_\sigma \models \phi$ iff $\phi \in \Delta_\sigma$.

Proof. As usual, by induction on $\phi$. 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 $M^\Gamma$ is indeed a game model. We start by looking at the epistemic relation $\sim^\Gamma_i$. 
5.5.21. **Lemma (Adequacy of $\sim^\Gamma_i$ - Part I).** The relation $\sim^\Gamma_i$ 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 $W^\Gamma$, for each agent. We can look at these partitions directly as strategies. That is, for each “profile” $\Delta_\sigma$, 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 $\sim^\Gamma_i$ is then easy.

5.5.22. **Lemma (Adequacy of $\sim^\Gamma_i$ - Part II).** For all $\Delta_\sigma$ and $\Delta'_\sigma$, if $\Delta_\sigma \sim^\Gamma_i \Delta'_\sigma$ then $\Delta_\sigma(i) = \Delta'_\sigma(i)$ and $i^\Gamma_\sigma(\Delta_\sigma) = i^\Gamma_\sigma(\Delta'_\sigma)$.

**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 $i^\Gamma_\sigma$ all we need to show is that for all $|\phi| \in i^\Gamma_\sigma(\Delta_\sigma)$, $|\phi|$ is also in $i^\Gamma_\sigma(\Delta'_\sigma)$. So assume the first. This means that $I_i \phi \in \Delta_\sigma$, which means by (K-I) that $K_i I_i \phi$ is also in $\Delta_\sigma$. But then, because $\Delta_\sigma \sim^\Gamma_i \Delta'_\sigma$, we obtain by a routine modal logic argument that $I_i \phi \in \Delta'_\sigma$, which is just to say, $|\phi|$ is also in $i^\Gamma_\sigma(\Delta'_\sigma)$.

5.5.23. **Lemma (Adequacy of $\geq^\Gamma_i$ and $\succ^\Gamma_i$).** The relation $\geq^\Gamma_i$ is a total, reflexive and transitive relation on $W^\Gamma$, and $\succ^\Gamma_i$ is its irreflexive and transitive sub-relation.

**Proof.** The $S4$ axioms for $\Diamond^\leq_i$ are canonical. Irreflexivity of $\succ_i$ and totality of $\geq_i$ 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-relation of $\geq_i$.

5.5.24. **Lemma (Adequacy of $i^\Gamma_i$).** For all $\Delta_\sigma$, $i^\Gamma_\sigma(\Delta_\sigma)$ 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 $DL_{\mathcal{G}_\mathcal{F}}$**

I define $\Lambda_{DL_{\mathcal{G}_\mathcal{F}}}$ as $\Lambda_{\mathcal{L}_{\mathcal{G}_\mathcal{F}}}$ 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.
Chapter 5. Logics for practical reasoning with intentions

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 \(D\mathcal{L}\bar{\phi}\)] Given a game model \(M\) with two agents, if at a profile \(w\) named by \(\sigma\),

\[M, w \models \sigma(2)!WR_1 \land \sigma(1)!WR_2\]

then \(w\) is a Nash equilibrium.
5.5. Appendix

Proof. The argument again boils down to showing that at \( w \) both agents play a best response. Consider player 1. Clearly \( M, w \models \sigma(2) \), and so it must be that \( M \models w \models \sigma(2) \). Now observe that \( W_{\sigma(2)} = \{ w' : w' = w[s/w(i)] \text{ for a } s \in S_i \} \). But this means, by the same argument as in Fact 5.2.12, that \( w \models_i \sigma(2) \). So \( M \models \sigma(2) \), \( w \models W R_1 \) 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.