

Axiomatizing Rationality

Adam Bjorn Dahl

Dept. of Mathematics
Cornell University
Ithaca, NY 14853
abjorn Dahl@math.cornell.edu

Joseph Y. Halpern

Dept. of Computer Science
Cornell University
Ithaca, NY 14853
halpern@cs.cornell.edu

Rafael Pass

Dept. of Computer Science
Cornell University
Ithaca, NY 14853
rafael@cs.cornell.edu

Abstract

We provide a sound and complete axiomatization for a class of logics appropriate for reasoning about the rationality of players in games. Essentially the same axiomatization applies to a wide class of decision rules.

1 Introduction

Decision-making in an uncertain environment is a fundamental component of game theory: players must choose what to do without necessarily knowing what their opponents will do. Under certainty, decision-making is straightforward; one simply chooses the course of action that leads to the most preferred outcome. Under uncertainty, however, a player must evaluate many possible outcomes in a manner that somehow takes into account her relative degrees of belief. There are many ways to do this. The *maximin* decision rule, for example, focuses entirely on worst-case scenarios. An *expected utility maximizer*, on the other hand, weights each outcome according to her (subjective) probability and chooses the course of action that maximizes the corresponding expected value.

One can argue about which decision rules are reasonable and which are not, and the word “rational” is often invoked to denote this very divide, but this is not the notion of rationality that concerns us here. Rather, assuming that we have fixed a decision rule for a given player, we can ask whether that player is, in fact, making choices in accordance with it, and call her *rational* precisely when she is. In classical game theory, for example, rationality is typically identified with expected utility maximization: a player is rational precisely if she is acting to maximize her expected utility.

Another fundamental component of game theory is higher-order belief: what you believe your opponent will do is connected with what you believe your opponent believes. In other words, your beliefs about their beliefs matter. Many notions of equilibrium, for example, require not only that each player is rational, but also that each player believes their opponents to be rational as well. Reasoning about games therefore requires reasoning about beliefs.

Epistemic game theory takes this as its motto, and concerns itself with models expressive enough to represent the

complexities of higher-order beliefs in a game-theoretic context. Formal logic furnishes a powerful and versatile class of such models; namely, modal logics of belief appropriately specialized for reasoning about games. Rationality is no less important in this setting; however, while it has been incorporated into these models both syntactically and semantically, no axiomatization of the resulting logical systems has been provided. This paper fills this gap.

We take as our point of departure axioms for rationality in the sense of expected utility maximization given in (Bjorn Dahl, Halpern, & Pass 2011). We extend these axioms to arbitrary decision rules, under the assumption that the players’ uncertainty is represented by a probability measure. This allows us to deal with not just expected utility maximization, but other standard rules such as *maximin* and *minimax regret* (see (Halpern 2003) for a discussion of all the decision rules mentioned in this paper). We then go on to consider what happens when the players’ uncertainty is represented by a set of probabilities, which allows us to capture well-known decision rules such as *maximin expected utility* and *minimax expected regret*. Finally, we consider situations where a player might be uncertain about which decision rules his opponents are using.

The rest of this paper is organized as follows. In Section 2, we define the core concepts formally: games, modal logics of belief appropriate for reasoning about games, and the incorporation of rationality into these logics. Section 3 gives the axiomatization, and proves that it is sound and complete. In Section 4, we provide sound and complete axiomatizations for the cases where the players’ uncertainty is represented by sets of probabilities, and where players may be uncertain about the decision rules used by other players. In Section 5 we discuss the role of language. Section 6 concludes with a discussion of future work. More detailed proofs and further discussion can be found in the full paper, which is available at http://www.math.cornell.edu/~abjorn Dahl/Site/CV_files/Axiomatizing%20Rationality.pdf.

2 Reasoning about games

Given a tuple $(X_i)_{i \in I}$ over some finite index set $I = \{1, \dots, n\}$, we adopt the usual notational convention of writing $X := \prod_{i \in I} X_i$ and $X_{-i} := \prod_{j \neq i} X_j$. We also

write $X'_i \times X_{-i}$ for

$$X_1 \times \cdots \times X_{i-1} \times X'_i \times X_{i+1} \times \cdots \times X_n$$

and similarly (x'_i, x_{-i}) for $(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)$.

A **(normal-form) game** is a tuple $\Gamma = (I, (\Sigma_i)_{i \in I}, (u_i)_{i \in I})$ where $I = \{1, \dots, n\}$ is the set of *players*, Σ_i is the (finite) set of *strategies available to player i* , and $u_i : \Sigma \rightarrow \mathbb{R}$ is *player i 's utility function*, where $\Sigma = \prod_i \Sigma_i$ denotes the set of *strategy profiles*.

2.1 Syntax

One way of reasoning formally about a game is to build a logical language that is expressive enough to capture the aspects of play that we are interested in analyzing. To this end, given a game Γ , we begin by defining a propositional modal language of belief and then specializing the primitive propositions to correspond to the strategies available to the players.

Given an arbitrary set Φ of primitive propositions, let $\mathcal{L}_B(\Phi)$ be the language recursively generated by the grammar

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid B_i\varphi,$$

where $p \in \Phi$ and $i \in I$. As usual, we read $B_i\varphi$ as “player i believes φ ”. We define $\vee$, $\rightarrow$, and $\leftrightarrow$ as usual in terms of $\wedge$ and $\neg$, and write $\hat{B}_i\varphi$ for $\neg B_i\neg\varphi$ (“player i consider it possible that φ ”). We also write

$$\begin{aligned} EB^1\varphi &\equiv B_1\varphi \wedge \cdots \wedge B_n\varphi \text{ and} \\ EB^k\varphi &\equiv EB^1(EB^{k-1}\varphi) \end{aligned}$$

for “everyone believes φ ” and its k -fold iteration, respectively. Set

$$\Phi_\Gamma := \{play_i(\sigma_i) : i \in I, \sigma_i \in \Sigma_i\},$$

where we read $play_i(\sigma_i)$ as “player i is playing strategy σ_i ”; we write

$$\begin{aligned} play(\sigma) &\equiv play_1(\sigma_1) \wedge \cdots \wedge play_n(\sigma_n) \text{ and} \\ play_{-i}(\sigma_{-i}) &\equiv \bigwedge_{j \neq i} play_j(\sigma_j) \end{aligned}$$

for “the players are playing according to the strategy profile σ ” and the analogous statement regarding the players other than i . The language $\mathcal{L}_B(\Phi_\Gamma)$ is thus appropriate for reasoning about the beliefs of the players with respect to the strategies they are playing.

2.2 Semantics

A language of belief can be interpreted using Kripke-style possible world semantics, where associated to each world ω and each player i is a probability measure on the set of all worlds, used to interpret player i 's beliefs at ω . In the case of a language like $\mathcal{L}_B(\Phi_\Gamma)$, we also must take care to interpret the primitive propositions appropriately.

We restrict our attention to finite models. A **finite Γ -structure** is a tuple $M = (\Omega, (Pr_i)_{i \in I}, s)$ satisfying the following conditions:

(C1) Ω is a nonempty, finite set;

(C2) Pr_i associates with each $\omega \in \Omega$ a probability measure $Pr_i(\omega)$ on Ω ;

(C3) $Pr_i(\omega)(\{\omega' \in \Omega : Pr_i(\omega') = Pr_i(\omega)\}) = 1$;

(C4) $s : \Omega \rightarrow \Sigma$ satisfies $Pr_i(\omega)(\{\omega' \in \Omega : s_i(\omega') = s_i(\omega)\}) = 1$.

Conditions (C1) and (C2) set the stage to interpret player i 's beliefs at ω by the measure $Pr_i(\omega)$. Condition (C3) then ensures that at each world ω , each player is sure of (i.e. assigns probability 1 to) her own beliefs. Finally, condition (C4) establishes that the *strategy function* s assigns to each world ω a strategy profile $s(\omega)$ in game Γ —intuitively, the strategy that each player is playing at ω —and moreover, that each player is sure of her own strategy.

A finite Γ -structure M induces an **interpretation** $\llbracket \cdot \rrbracket_M : \mathcal{L}_B(\Phi_\Gamma) \rightarrow 2^\Omega$ defined recursively as follows:

$$\begin{aligned} \llbracket play_i(\sigma_i) \rrbracket_M &:= \{\omega \in \Omega : s_i(\omega) = \sigma_i\} \\ \llbracket \varphi \wedge \psi \rrbracket_M &:= \llbracket \varphi \rrbracket_M \cap \llbracket \psi \rrbracket_M \\ \llbracket \neg\varphi \rrbracket_M &:= \Omega \setminus \llbracket \varphi \rrbracket_M \\ \llbracket B_i\varphi \rrbracket_M &:= \{\omega \in \Omega : Pr_i(\omega)(\llbracket \varphi \rrbracket_M) = 1\}. \end{aligned}$$

Thus, the primitive propositions are interpreted in the obvious way using the strategy function (s_i denotes the i th component function of s), the Boolean connectives are interpreted classically, and the formula $B_i\varphi$ holds at all and only those worlds ω at which $Pr_i(\omega)$ assigns probability 1 to φ . As is standard, we often write $(M, \omega) \models \varphi$ or just $\omega \models \varphi$ to indicate that $\omega \in \llbracket \varphi \rrbracket_M$. Similarly, we write $M \models \varphi$ if $\llbracket \varphi \rrbracket_M = \Omega$, and when $(M, \omega) \not\models \varphi$, we say that M *refutes* φ (at ω), or just that ω refutes φ .

2.3 Rationality

Informally, a player is *rational* if the strategy she is playing is a best response to her beliefs about the outcome of the game, given her preferences. But there is no single conception of what constitutes a “best response”; a variety of principles of decision-making have been proposed and studied.

One influential notion, particularly in game theory, is that of expected utility maximization. Given a game Γ and a probability measure μ on Σ_{-i} (thought of as representing player i 's beliefs about the strategies her opponents will play), the **expected utility** of a strategy $\sigma_i \in \Sigma_i$ is just the expected value of the function $u_i(\sigma_i, \cdot) : \Sigma_{-i} \rightarrow \mathbb{R}$ with respect to μ :

$$EU_i(\sigma_i; \mu) := \sum_{\sigma_{-i} \in \Sigma_{-i}} u_i(\sigma_i, \sigma_{-i}) \cdot \mu(\sigma_{-i}).$$

A best response for an expected utility maximizer is a strategy that maximizes this value. Abstractly, we might identify the mandate “maximize expected utility” for player i (in the game Γ) with a function $\mathfrak{d}_i^{eu}$ that takes as input a belief μ on Σ_{-i} and returns as output the set of strategies $\sigma_i \in \Sigma_i$ that maximize player i 's expected utility given μ :

$$\mathfrak{d}_i^{eu}(\mu) := \{\sigma_i \in \Sigma_i : (\forall \sigma'_i \in \Sigma_i)(EU_i(\sigma_i; \mu) \geq EU_i(\sigma'_i; \mu))\}.$$

Let $\Delta(\Sigma_{-i})$ denote the set of all probability measures on Σ_{-i} . Then, generalizing the above, we define a **decision**

rule for player i (in Γ) to be a function $\mathfrak{d}_i : \Delta(\Sigma_{-i}) \rightarrow 2^{\Sigma_i} \setminus \{\emptyset\}$. Intuitively, $\sigma_i \in \mathfrak{d}_i(\mu)$ just in case σ_i is a best response for player i to the belief μ according to the decision rule $\mathfrak{d}_i$. For example, the “maximin” mandate for player i , which says to maximize the worst-case outcome among those considered possible, corresponds to the decision rule defined as follows: let

$$WC_i(\sigma_i; \mu) := \min\{u_i(\sigma_i, \sigma_{-i}) : \mu(\sigma_{-i}) > 0\},$$

and set

$$\mathfrak{d}_i^m(\mu) := \{\sigma_i \in \Sigma_i : (\forall \sigma'_i \in \Sigma_i)(WC_i(\sigma_i; \mu) \geq WC_i(\sigma'_i; \mu))\}.$$

Decision rules can be interpreted in Γ -structures; roughly speaking, for each world ω , we can define the set of $\mathfrak{d}_i$ -best responses for player i at ω , and thus determine whether or not player i is being $\mathfrak{d}_i$ -rational (i.e., acting in accordance with the decision rule $\mathfrak{d}_i$) at ω . Formally, given a Γ -structure M , for each player i and each world ω , the probability measure $Pr_i(\omega)$ induces a probability measure $\mu_{i,\omega}$ defined on Σ_{-i} as follows:

$$\mu_{i,\omega}(\sigma_{-i}) := Pr_i(\omega)(\llbracket play_{-i}(\sigma_{-i}) \rrbracket_M).$$

In fact, $\mu_{i,\omega}$ is the *pushforward* of $Pr_i(\omega)$ by s_{-i} . Since $Pr_i(\omega)$ is interpreted as representing player i 's beliefs at ω , it makes sense to apply her decision rule to $\mu_{i,\omega}$. This leads us to define the set of **$\mathfrak{d}_i$ -best responses for player i at ω** to be $\mathfrak{d}_i(\mu_{i,\omega})$. We say that player i is **$\mathfrak{d}_i$ -rational at ω** just in case $s_i(\omega) \in \mathfrak{d}_i(\mu_{i,\omega})$.

Since we wish to reason formally about games using the logic developed above, and rationality is an important concept for such analyses, we expand the language to include primitive propositions denoting rationality of the players. Given a profile of decision rules $\mathfrak{d} = (\mathfrak{d}_i)_{i \in I}$, let

$$\Phi_\Gamma^\mathfrak{d} := \Phi_\Gamma \cup \{RAT_i^{\mathfrak{d}_i} : i \in I\},$$

where $RAT_i^{\mathfrak{d}_i}$ is read “player i is $\mathfrak{d}_i$ -rational”. Henceforth, except where noted otherwise, we work with an arbitrary but fixed profile of decision rules $\mathfrak{d}$. For notational convenience, we omit the $\mathfrak{d}_i$ when possible, referring to “best responses” and “rationality” instead of “ $\mathfrak{d}_i$ -best responses” and “ $\mathfrak{d}_i$ -rationality”, and writing RAT_i instead of $RAT_i^{\mathfrak{d}_i}$. We also make use of the syntactic abbreviation

$$RAT \equiv RAT_1 \wedge \dots \wedge RAT_n$$

for “everyone is rational”.

Given a Γ -structure M , the interpretation $\llbracket \cdot \rrbracket_M$ is extended to $\mathcal{L}_B(\Phi_\Gamma^\mathfrak{d})$ in the obvious way, by setting

$$\llbracket RAT_i \rrbracket_M := \{\omega \in \Omega : s_i(\omega) \in \mathfrak{d}_i(\mu_{i,\omega})\}.$$

When $\mathfrak{d} = \mathfrak{d}^{eu}$, so each player's decision rule is given by expected utility maximization, we regain the traditional notion of rationality in game theory. Rationality so defined can be used to characterize several well-known solution concepts in terms of Γ -structures. For example, as shown by Tan and Werlang (1988) and Brandenburger and Dekel (1987), given a game Γ , a strategy σ_i is *rationalizable* if and only if there exists a Γ -structure M and a state ω therein such that

$\omega \models play_i(\sigma_i)$ and for every $k \in \mathbb{N}$, $\omega \models EB^k(RAT)$ (i.e. it is common belief that everyone is rational).

It is worth noting that decision rules are general enough to represent processes that fall outside the traditional purview of “rationality”. For example, suppose that player i is a computer system and Σ_i is a collection of actions it can execute. Suppose also that the system maintains a database consisting of estimates of the values of certain variables, which can be represented as beliefs about the strategies of “opponents”. In this context, a decision rule can be thought of as a non-deterministic process that the computer system might use to choose which action to execute on the basis of the information in its database. In particular, $RAT_i^{\mathfrak{d}_i}$ asserts that the system has executed an action consistent with the nondeterministic process $\mathfrak{d}_i$.¹

2.4 Example

In order to help solidify the framework just introduced, we present a simple example. Consider the standard Bach-or-Stravinsky game Γ (Osborne & Rubinstein 1994), in which each of two players must choose which of two concerts to attend this evening: one featuring the music of Bach, and one of Stravinsky. Player 1 prefers to attend the Bach concert, while player two prefers the Stravinsky; however, each much prefers to attend the same concert as the other. We can represent these preferences with the utility functions summarized in Table 1.

	Bach	Stravinsky
Bach	3,2	0,0
Stravinsky	0,0	2,3

Table 1: Bach or Stravinsky?

We now describe a Γ -structure in which we can reason about the beliefs and rationality of the players. Let $\Omega = \{\omega_0, \omega_1, \omega_2\}$ and let $Pr_1(\omega_0)$ be the uniform distribution on $\{\omega_1, \omega_2\}$, so player 1 considers ω_1 and ω_2 equally likely. Consider a strategy function s defined such that $s_2(\omega_1) = \text{Bach}$ and $s_2(\omega_2) = \text{Stravinsky}$. It is easy to see that in this case, the induced measure μ_{1,ω_0} assigns probability .5 to Bach and .5 to Stravinsky. It follows easily that player 1's expected utility on choosing Bach is 1.5, while his expected utility on choosing Stravinsky is 1. Thus,

$$\omega_0 \models RAT_1^{\mathfrak{d}_1^{eu}} \leftrightarrow play_1(\text{Bach}).$$

On the other hand, the worst-case outcome for player 1 on choosing either Bach or Stravinsky yields a utility of 0, so we have

$$\omega_0 \models RAT_1^{\mathfrak{d}_1^m},$$

regardless of which strategy player 1 actually chooses at ω_0 .

¹In this sense, we can think of a decision rule as a *knowledge-based program* (Fagin *et al.* 1995). Knowledge-based programs have previously been used to characterize rationality and solution concepts (Halpern & Moses 2007).

3 Axiomatization

Although reasoning formally about beliefs and rationality has long been recognized as important to game theory (see, e.g., (?)), to the best of our knowledge, rationality has not been axiomatized in a logic of belief. In this section we provide an axiomatization and prove that it is sound and complete.

3.1 The axioms

While most of the axioms are easy to state and understand, two require some preliminary definitions. Fix a game Γ and a profile of decision rules $\mathfrak{d}$. Given a subset $S \subseteq \Sigma_{-i}$, let

$$\delta_{i,S} \equiv \bigwedge_{\sigma_{-i} \in S} \widehat{B}_i \text{play}_{-i}(\sigma_{-i}) \wedge \bigwedge_{\sigma_{-i} \notin S} \neg \widehat{B}_i \text{play}_{-i}(\sigma_{-i}).$$

Intuitively, the formula $\delta_{i,S}$ says that player i considers possible *all* and *only* the strategy profiles for his opponents that are elements of S . It is easy to see that for each player i and any world ω , exactly one of the formulas $\delta_{i,S}$ holds.

Given a measure μ on a finite set X , let

$$\text{supp}(\mu) := \{x \in X : \mu(x) > 0\},$$

the **support** of μ . For each player i and each $\sigma_i \in \Sigma_i$, let $\mathcal{S}_i^+(\sigma_i)$ denote the collection of all $S \subseteq \Sigma_{-i}$ such that there exists a probability measure μ with $\text{supp}(\mu) = S$ and with respect to which σ_i is a best response; that is, $\sigma_i \in \mathfrak{d}_i(\mu)$. Similarly, define $\mathcal{S}_i^-(\sigma_i)$ to be the collection of all $S \subseteq \Sigma_{-i}$ such that there exists a probability measure μ with $\text{supp}(\mu) = S$ and $\sigma_i \notin \mathfrak{d}_i(\mu)$.

Consider the following axiom schemes:

- G1.** $\bigvee_{\sigma_i \in \Sigma_i} \text{play}_i(\sigma_i)$
- G2.** $\neg(\text{play}_i(\sigma_i) \wedge \text{play}_i(\sigma'_i))$, for $\sigma_i \neq \sigma'_i$
- G3.** $\text{play}_i(\sigma_i) \leftrightarrow B_i \text{play}_i(\sigma_i)$
- G4.** $RAT_i \leftrightarrow B_i(RAT_i)$
- G5.** $(\text{play}_i(\sigma_i) \wedge RAT_i) \rightarrow \bigvee_{S \in \mathcal{S}_i^+(\sigma_i)} \delta_{i,S}$
- G6.** $(\text{play}_i(\sigma_i) \wedge \neg RAT_i) \rightarrow \bigvee_{S \in \mathcal{S}_i^-(\sigma_i)} \delta_{i,S}$.

G1–G4 are straightforward. **G1** and **G2** say that, at each state, each player plays exactly one strategy. **G3** and **G4** say that each player is certain of her strategy and of whether or not she is rational. The interesting axioms are **G5** and **G6**. Intuitively, **G5** says that if RAT_i holds and player i is playing σ_i , then player i must consider possible a collection of strategy profiles² on which she could place a probability that would justify her playing σ_i . **G6** is interpreted analogously. Notice that player i 's actual (i.e. quantitative)

²Note that the sets $\mathcal{S}_i^+(\sigma_i)$ and $\mathcal{S}_i^-(\sigma_i)$ could be empty; by convention, we define the empty disjunction to be $\perp$ (or *false*). If, for example, $\mathcal{S}_i^+(\sigma_i) = \emptyset$, this means that σ_i is not a best response to any beliefs, in which case the corresponding axiom $(\text{play}_i(\sigma_i) \wedge RAT_i) \rightarrow \perp$ is intuitively correct.

beliefs are not fully specified. This may be somewhat surprising, given that the semantic interpretation of RAT_i is defined in terms of the quantitative probabilities that constitute player i 's beliefs. Nevertheless, as we show in Section 3.2, these axioms completely characterize the logical properties of rationality as expressible in the language $\mathcal{L}_B(\Phi_\Gamma^\mathfrak{d})$; we return to this point in Section 5. Axioms **G1–G4** and the specialization of axioms **G5** and **G6** to the rule of maximizing expected utility appear in (Bjorndahl, Halpern, & Pass 2011).

Let $GL_\Gamma^\mathfrak{d}$ be the axiom system that results from adding **G1–G6** to the standard KD45 axioms and rules of inference of belief logic (see, e.g., (Fagin *et al.* 1995)). It bears emphasizing that $GL_\Gamma^\mathfrak{d}$ is parametrized by two variables: the underlying game Γ that determines the players, their strategies, and their preferences, and the profile of decision rules $\mathfrak{d}$ that determines the meaning of “rationality” for each player. Thus, what we are axiomatizing here is not a single logic but a class of logics.

For each fixed Γ and $\mathfrak{d}$, the corresponding axiom system $GL_\Gamma^\mathfrak{d}$ is trivially decidable (i.e., we can effectively determine whether a formula is an instance of an axiom scheme) because KD45 is decidable and there are only finitely-many instances of the axiom schemes **G1–G6** (the implicit quantification in these schemes ranges in some cases over players and in others over strategies, all of which are finite sets).

But there is arguably a more interesting question. Up to now we have fixed a game Γ , and considered decision rules in the game Γ . But rules like maximizing expected utility, maximin, or minimizing regret can be applied in all games in a uniform way. To capture this, we define a **decision paradigm** to be a function that maps each game Γ to a decision rule in Γ . Suppose that we are given decision paradigms $\mathcal{D}_i$ for each player associating with each game Γ a decision rule $\mathcal{D}_i(\Gamma)$ for that player in Γ . We might want to know, given the profile $\mathcal{D} = (\mathcal{D}_i)_{i \in I}$, whether the mapping

$$\Gamma \mapsto GL_\Gamma^{\mathcal{D}(\Gamma)}$$

is decidable; in other words, given as input a game Γ , can we effectively determine whether a formula belongs to the axiom system $GL_\Gamma^{\mathcal{D}(\Gamma)}$? For each game Γ , this requires determining membership in the sets $\mathcal{S}_i^+(\sigma_i)$ and $\mathcal{S}_i^-(\sigma_i)$, which are defined by existential quantification over simplices $\Delta(\Sigma_{-i})$, subject to constraints based on the decision rules $\mathcal{D}_i(\Gamma)$. In the case of familiar decision paradigms like maximin or expected utility maximization, computing the sets $\mathcal{S}_i^+(\sigma_i)$ and $\mathcal{S}_i^-(\sigma_i)$ boils down to solving systems of linear inequalities. In general, however, we must impose certain computability requirements on the decision paradigms in order to be able to decide whether a formula is an instance of an axiom. At this point, we have not found an interesting characterization of when we get decidability.

3.2 Soundness and completeness

Theorem 3.1: $GL_\Gamma^\mathfrak{d}$ is a sound axiomatization of the language $\mathcal{L}_B(\Phi_\Gamma^\mathfrak{d})$ with respect to the class of all finite Γ -structures.

Proof: The proof is much in the same spirit as the proof of the corresponding result in (Bjorndahl, Halpern, & Pass 2011) (except that now we consider rules more general than just maximizing expected utility). Soundness of the axioms and rules of KD45 can be proved as usual. It therefore suffices to show that **G1**–**G6** are valid in all finite Γ -structures.

Fix a finite Γ -structure $M = (\Omega, (Pr_i)_{i \in I}, s)$. Soundness of **G1** and **G2** is an immediate consequence of the fact that s is a (total) function. Soundness of **G3** is a straightforward consequence of condition (C4), while soundness of **G4** follows easily from the combination of conditions (C3) and (C4).

Now suppose that $\omega \models play_i(\sigma_i) \wedge RAT_i$; then $\sigma_i \in \mathfrak{d}_i(\mu_{i,\omega})$. Set $S := \text{supp}(\mu_{i,\omega})$, and observe that $S \in \mathcal{S}_i^+(\sigma_i)$. For each $\sigma_{-i} \in S$, it is easy to see that $\omega \models \widehat{B}_i play_{-i}(\sigma_{-i})$; therefore, we must have

$$\omega \models \bigwedge_{\sigma_{-i} \in S} \widehat{B}_i play_{-i}(\sigma_{-i}).$$

Similarly, for each $\sigma_{-i} \notin S$, we have $\omega \models \neg \widehat{B}_i play_{-i}(\sigma_{-i})$, so

$$\omega \models \bigwedge_{\sigma_{-i} \notin S} \neg \widehat{B}_i play_{-i}(\sigma_{-i}).$$

In other words, $\omega \models \delta_{i,S}$; this establishes the soundness of **G5**.

Finally, suppose that $\omega \models play_i(\sigma_i) \wedge \neg RAT_i$, which implies that $\sigma_i \notin \mathfrak{d}_i(\mu_{i,\omega})$. Set $S := \text{supp}(\mu_{i,\omega})$; then $S \in \mathcal{S}_i^-(\sigma_i)$ and, as above, we have $\omega \models \delta_{i,S}$, which establishes soundness of **G6**. ■

We prove completeness by what is essentially the *canonical model method*, a standard method for proving completeness of modal logics (see, e.g., (Fagin *et al.* 1995) or any standard text on modal logic). Of course, the full canonical model is not finite (for $n > 1$), so we modify the construction by restricting attention to finite sub-languages. More precisely, given a formula $\varphi \in \mathcal{L}_B(\Phi_\Gamma^0)$, we identify a finite sub-language of $\mathcal{L}_B(\Phi_\Gamma^0)$ such that the corresponding canonical model refutes φ just in case $GL_\Gamma^0 \not\models \varphi$. This technique is sometimes called *filtration*.

Fix a formula $\varphi \in \mathcal{L}_B(\Phi_\Gamma^0)$. Let $Sub_\Gamma(\varphi)$ denote the collection of all subformulas of φ together with all subformulas of instances of the axiom schemes **G1** through **G6**. Define

$$Sub_\Gamma^+(\varphi) := Sub_\Gamma(\varphi) \cup \{\neg\psi : \psi \in Sub_\Gamma(\varphi)\}.$$

Note that there are only finitely many instances of **G1** through **G6**, and therefore $Sub_\Gamma^+(\varphi)$ is finite.

Let Ω^φ be the collection of all maximal, consistent (with respect to GL_Γ^0) subsets of $Sub_\Gamma^+(\varphi)$. Clearly Ω^φ is a finite set. Given $X \subseteq \mathcal{L}_B(\Phi_\Gamma^0)$, set

$$X^{B_i} := \{\psi : B_i\psi \in X\};$$

and, for $F \in \Omega^\varphi$, define

$$Bel_i(F) := \{G \in \Omega^\varphi : G \supseteq F^{B_i} \text{ and } G^{B_i} = F^{B_i}\}.$$

For each $i \in I$ and each $F \in \Omega^\varphi$, we will define a probability measure $Pr_i^\varphi(F)$ on Ω^φ such that the support of this

measure is precisely $Bel_i(F)$. Loosely speaking, $Bel_i(F)$ is the set of all $G \in \Omega^\varphi$ that are compatible with the beliefs of player i in F . More precisely, $G \in Bel_i(F)$ if and only if:

- (a) $B_i\psi \in F$ implies $\psi \in G$, and
- (b) $B_i\psi \in F$ iff $B_i\psi \in G$.

Condition (a) just says that everything player i believes in F is true in G ; condition (b) says that player i 's beliefs in F are the same as her beliefs in G , which is reasonable in light of the fact that we are working in a system with positive and negative introspection. (In the full canonical model, (b) follows from (a); here we must impose this condition explicitly because of the way the language has been restricted.)

Since our aim is to define probability measures with the sets $Bel_i(F)$ as their supports, we must show that these sets are never empty.

Lemma 3.2: For each $i \in I$ and $F \in \Omega^\varphi$, $Bel_i(F) \neq \emptyset$.

Proof: Given $F \in \Omega^\varphi$, set

$$\Lambda := \{\psi : B_i\psi \in F\} \cup \{B_i\psi : B_i\psi \in F\} \\ \cup \{\neg B_i\psi : \neg B_i\psi \in F\}.$$

It is easy to see that $\Lambda \subset Sub_\Gamma^+(\varphi)$. In addition, we show that Λ is consistent. For suppose not; then

$$\begin{aligned} GL_\Gamma^0 \vdash \neg \bigwedge_{\xi \in \Lambda} \xi &\implies GL_\Gamma^0 \vdash B_i \neg \bigwedge_{\xi \in \Lambda} \xi \\ &\implies GL_\Gamma^0 \vdash \neg B_i \bigwedge_{\xi \in \Lambda} \xi \\ &\implies GL_\Gamma^0 \vdash \neg \bigwedge_{\xi \in \Lambda} B_i \xi, \end{aligned}$$

which is a contradiction, since each formula $B_i\xi$ with $\xi \in \Lambda$ is logically equivalent to a formula in F , and F is consistent.

From this we can conclude that there exists a $G \in \Omega^\varphi$ such that $G \supseteq \Lambda$. It follows immediately that $G \supseteq F^{B_i}$ and that $G^{B_i} \supseteq F^{B_i}$. Moreover, if $\psi \in G^{B_i}$, then $B_i\psi \in G$ and so certainly $\neg B_i\psi \notin G$, from which it follows that $\neg B_i\psi \notin \Lambda$ and thus $\neg B_i\psi \notin F$. Maximality of F then guarantees that $B_i\psi \in F$, whence $\psi \in F^{B_i}$, and so $G^{B_i} \subseteq F^{B_i}$. This establishes that $G \in Bel_i(F)$, as desired. ■

In the classical canonical model construction, it is sufficient to define $Pr_i^\varphi(F)$ to be the uniform distribution on $Bel_i(F)$. In the present context, however, we need to be more careful, since $Pr_i^\varphi(F)$ is used not only to interpret the belief modalities B_i , but also the primitive propositions RAT_i . In essence, we must define $Pr_i^\varphi(F)$ in a manner that agrees with whether or not player i is best responding to her beliefs at F ; not surprisingly, this is precisely where the axiom schemes **G5** and **G6** come into play. At the same time, we have to define Pr_i^φ on Ω^φ in a systematic way so as to preserve the introspection condition (C3). What follows is a formalization of this basic recipe, for which several more lemmas and definitions are needed.

Lemma 3.3: Let $F \in \Omega^\varphi$. If $\widehat{B}_i\psi \in F$, then there exists a $G \in Bel_i(F)$ with $\psi \in G$; if $\neg \widehat{B}_i\psi \in F$, then for all $G \in Bel_i(F)$ we have $\psi \notin G$.

Proof: See the full paper. ■

For each $\sigma_{-i} \in \Sigma_{-i}$, define

$$\text{Bel}_i(F; \sigma_{-i}) := \{G \in \text{Bel}_i(F) : \text{play}_{-i}(\sigma_{-i}) \in G\}.$$

Given $F \in \Omega^\varphi$, it is easy to see, using **G1** and **G2**, that there is a unique $\sigma_i \in \Sigma_i$ with $\text{play}_i(\sigma_i) \in F$. If, in addition, $\text{RAT}_i \in F$, then by **G5** we know that for some $S \in \mathcal{S}_i^+(\sigma_i)$, $\delta_{i,S} \in F$ (or, in the case where $\mathcal{S}_i^+(\sigma_i) = \emptyset$, we know that no such F exists). Otherwise, if $\text{RAT}_i \notin F$, then by **G6** we know that for some $S \in \mathcal{S}_i^-(\sigma_i)$, $\delta_{i,S} \in F$ (or again, in the case where $\mathcal{S}_i^-(\sigma_i) = \emptyset$, that no such F exists). Thus, for each $i \in I$, there is a unique set $S_i(F) \subseteq \Sigma_{-i}$ such that $\delta_{i,S_i(F)} \in F$, and moreover, $S_i(F) \in \mathcal{S}_i^+(\sigma_i)$ if $\text{RAT}_i \in F$, and $S_i(F) \in \mathcal{S}_i^-(\sigma_i)$ if $\text{RAT}_i \notin F$.

Lemma 3.4: *The collection $\{\text{Bel}_i(F; \sigma_{-i}) : \sigma_{-i} \in \Sigma_{-i}\}$ partitions $\text{Bel}_i(F)$; moreover, $\text{Bel}_i(F; \sigma_{-i}) \neq \emptyset$ if and only if $\sigma_{-i} \in S_i(F)$.*

Proof: The first statement is a straightforward consequence of **G1** and **G2**, while the second is an immediate corollary of Lemma 3.3 together with the fact that

$$\delta_{i,S_i(F)} \equiv \bigwedge_{\sigma_{-i} \in S_i(F)} \hat{B}_i \text{play}_{-i}(\sigma_{-i}) \wedge \bigwedge_{\sigma_{-i} \notin S_i(F)} \neg \hat{B}_i \text{play}_{-i}(\sigma_{-i}) \in F. \blacksquare$$

For each $\sigma_i \in \Sigma_i$ and $S \in \mathcal{S}_i^+(\sigma_i)$, let $\mu_{\sigma_i, S}^+$ be a fixed probability measure witnessing the fact that $S \in \mathcal{S}_i^+(\sigma_i)$; that is, $\text{supp}(\mu_{\sigma_i, S}^+) = S$ and $\sigma_i \in \mathfrak{d}_i(\mu_{\sigma_i, S}^+)$. Likewise, for each $\sigma_i \in \Sigma_i$ and $S \in \mathcal{S}_i^-(\sigma_i)$, let $\mu_{\sigma_i, S}^-$ be a fixed probability measure witnessing the fact that $S \in \mathcal{S}_i^-(\sigma_i)$.

Let $F \in \Omega^\varphi$, and suppose that $\text{play}_i(\sigma_i) \in F$. In light of Lemma 3.4, we can define $\text{Pr}_i^\varphi(F)$ to be the unique probability measure on $\text{Bel}_i(F)$ such that, for all $\sigma_{-i} \in \Sigma_{-i}$,

$$\text{Pr}_i^\varphi(F)(\text{Bel}_i(F; \sigma_{-i})) = \begin{cases} \mu_{\sigma_i, S_i(F)}^+(\sigma_{-i}) & \text{if } \text{RAT}_i \in F \\ \mu_{\sigma_i, S_i(F)}^-(\sigma_{-i}) & \text{if } \text{RAT}_i \notin F, \end{cases}$$

and which is uniform within each (nonempty) set $\text{Bel}_i(F; \sigma_{-i})$.

Proposition 3.5: *Pr_i^φ satisfies the following:*

- (a) $\text{Pr}_i^\varphi(F)(G) > 0$ iff $G \in \text{Bel}_i(F)$, and
- (b) $G^{B_i} = F^{B_i}$ implies $\text{Pr}_i^\varphi(G) = \text{Pr}_i^\varphi(F)$.

Proof: See the full paper. ■

Finally, we define a strategy function $s^\varphi : \Omega^\varphi \rightarrow \Sigma$ by assigning to each $F \in \Omega^\varphi$ the unique strategy profile $\sigma \in \Sigma$ such that $\text{play}(\sigma) \in F$.

Lemma 3.6: *The tuple $M^\varphi := (\Omega^\varphi, (\text{Pr}_i^\varphi)_{i \in I}, s^\varphi)$ is a Γ -structure.*

Proof: Conditions (C1) and (C2) have already been established. By Lemma 3.5(b), in order to see that (C3) holds it suffices to observe that $\text{Pr}_i^\varphi(F)(G) > 0$ implies that $G^{B_i} = F^{B_i}$, which follows from Lemma 3.5(a). This same observation also establishes (C4), since by **G3** we know that $G^{B_i} = F^{B_i}$ implies $s_i^\varphi(G) = s_i^\varphi(F)$. ■

Lemma 3.7: *For all formulas ψ , for all $F \in \Omega^\varphi$, if $\psi \in \text{Sub}_\Gamma^+(\varphi)$ then $F \in \llbracket \psi \rrbracket_{M^\varphi}$ if and only if $\psi \in F$.*

Proof: The proof proceeds by induction on the structure of ψ . We prove here the base cases corresponding to the primitive propositions; the inductive steps can be proved in the standard way (see, e.g., (Fagin *et al.* 1995)).

First consider the primitive proposition $\text{play}_i(\sigma_i)$. We have

$$\begin{aligned} F \in \llbracket \text{play}_i(\sigma_i) \rrbracket_{M^\varphi} & \text{ iff } s_i^\varphi(F) = \sigma_i \\ & \text{ iff } \text{play}_i(\sigma_i) \in F, \end{aligned}$$

as a direct consequence of the definition of s^φ . Next consider the primitive proposition RAT_i ; we have

$$\begin{aligned} F \in \llbracket \text{RAT}_i \rrbracket_{M^\varphi} & \text{ iff } s_i^\varphi(F) \in \mathfrak{d}_i(\mu_{i,F}) \\ & \text{ iff } \text{RAT}_i \in F, \end{aligned}$$

the last equivalence being a consequence of the definition of Pr_i^φ , which ensures that $s_i^\varphi(F)$ is a best response to (the pushforward of) $\text{Pr}_i^\varphi(F)$ precisely when $\text{RAT}_i \in F$. This completes the proof. ■

Theorem 3.8: *$\text{GL}_\Gamma^\varnothing$ is a complete axiomatization of the language $\mathcal{L}_B(\Phi_\Gamma^\varnothing)$ with respect to the class of all finite Γ -structures.*

Proof: Suppose that $\text{GL}_\Gamma^\varnothing \not\models \varphi$. Then $\{\neg\varphi\}$ is consistent and so can be extended to a maximal consistent set $F \in \Omega^\varphi$. By Lemma 3.7, this implies that $F \notin \llbracket \varphi \rrbracket_{M^\varphi}$ and so, in particular, that $M^\varphi \not\models \varphi$, as desired. ■

4 Extensions

4.1 Belief as lower probability

In this paper we have taken for granted that each player's uncertainty is represented by a probability measure. While this is a very standard assumption, it is by no means the only framework that has been considered; see (Halpern 2003) for an overview of different ways of modeling uncertainty. Here we show that, with very minor modifications, the results of Section 3.2 hold also in the more general context where beliefs are represented using *sets* of probability measures.

Given a set $\mathcal{P}$ of probability measures, the **lower probability** of an event E , denoted $\mathcal{P}_*(E)$, is defined to be the infimum of the probabilities assigned to E by members of $\mathcal{P}$:

$$\mathcal{P}_*(E) := \inf\{\mu(E) : \mu \in \mathcal{P}\}.$$

Fix a game Γ . A **finite lower Γ -structure** is a tuple $M = (\Omega, (\mathcal{P}_i)_{i \in I}, s)$ satisfying the following conditions:

- (L1) Ω is a nonempty, finite set;
- (L2) $\mathcal{P}_i$ associates to each $\omega \in \Omega$ a set $\mathcal{P}_i(\omega)$ of probability measures on Ω ;
- (L3) $\mathcal{P}_i(\omega)_*(\{\omega' \in \Omega : \mathcal{P}_i(\omega') = \mathcal{P}_i(\omega)\}) = 1$;
- (L4) $s : \Omega \rightarrow \Sigma$ satisfies $\mathcal{P}_i(\omega)_*(\{\omega' \in \Omega : s_i(\omega') = s_i(\omega)\}) = 1$.

These conditions are simply the analogues of conditions (C1) through (C4) where uncertainty is represented by sets of probability measures and certainty is identified with lower probability 1. Accordingly, we define the interpretation $\llbracket \cdot \rrbracket_M$ as before, except for the clause corresponding to the belief modalities, which is replaced by the following:

$$\llbracket B_i \varphi \rrbracket_M := \{\omega \in \Omega : \mathcal{PR}_i(\omega)_*(\llbracket \varphi \rrbracket_M) = 1\}.$$

Finally, a **decision rule** for player i in this context is a function $\mathfrak{d}_i : 2^{\Delta(\Sigma_{-i})} \rightarrow 2^{\Sigma_i}$, since player i must make her choice based on the uncertainty given by a set of probability measures. For example, the “maximin expected utility” decision rule for player i would be given by the following:

$$\mathfrak{d}_i^{meu}(\mathcal{P}) := \{\sigma_i \in \Sigma_i : (\forall \sigma'_i \in \Sigma_i) (\min_{\mu \in \mathcal{P}} \{EU_i(\sigma_i; \mu)\} \geq \min_{\mu \in \mathcal{P}} \{EU_i(\sigma'_i; \mu)\})\}.$$

Other rules, such as minimax expected regret (Hayashi 2008), can also easily be defined in this setting.

As before, such decision rules makes sense in a Γ -structure M : for each player i and each world ω , the probability measures in the set $\mathcal{PR}_i(\omega)$ can be pushed forward by s_{-i} to probability measures on Σ_{-i} . Let $\mathcal{P}_{i,\omega}$ denote the set of all such pushforwards:

$$\mathcal{P}_{i,\omega} := \{\mu_{i,\omega} : \mu \in \mathcal{PR}_i(\omega)\}.$$

Then we can define **$\mathfrak{d}_i$ -rationality** for player i at ω in the obvious way, namely, by the requirement that $s_i(\omega) \in \mathfrak{d}_i(\mathcal{P}_{i,\omega})$.

The axiomatization of Section 3 can be generalized as well. First observe that the dual belief modality, $\widehat{B}_i \equiv \neg B_i \neg$, is interpreted as positive *upper* probability, where the **upper probability** of an event E with respect to a set $\mathcal{P}$ of probability measures, denoted $\mathcal{P}^*(E)$, is given by

$$\mathcal{P}^*(E) := \sup\{\mu(E) : \mu \in \mathcal{P}\}.$$

Accordingly, given a set $\mathcal{P}$ of probability measures on a finite space X , we define

$$\text{supp}(\mathcal{P}) := \{x \in X : (\exists \mu \in \mathcal{P})(\mu(x) > 0)\}.$$

For each $\sigma_i \in \Sigma_i$, let $\mathcal{S}_i^+(\sigma_i)$ denote the collection of all $S \subseteq \Sigma_{-i}$ such that there exists a set of probability measures $\mathcal{P}$ with $\text{supp}(\mathcal{P}) = S$ and $\sigma_i \in \mathfrak{d}_i(\mathcal{P})$. Similarly, define $\mathcal{S}_i^-(\sigma_i)$ to be the collection of all $S \subseteq \Sigma_{-i}$ such that there exists a set $\mathcal{P}$ of probability measures with $\text{supp}(\mathcal{P}) = S$ and $\sigma_i \notin \mathfrak{d}_i(\mathcal{P})$. It is not hard to see that all of these definitions generalize what was presented in Sections 2 and 3; indeed, by considering the special case where all sets of probability measures are singletons, we recover that framework exactly. Moreover, the axiom system $\text{GL}_\Gamma^\mathfrak{d}$, interpreted in this more general setting using the definitions above, remains sound and complete.

Theorem 4.1: $\text{GL}_\Gamma^\mathfrak{d}$ is a sound and complete axiomatization of the language $\mathcal{L}_B(\Phi_\Gamma^\mathfrak{d})$ with respect to the class of all finite lower Γ -structures.

Proof: The proof given in Section 3.2 works here as well, modulo the obvious minor alterations in keeping with the generalized definitions given above. In particular, for each

$\sigma_i \in \Sigma_i$ and $S \in \mathcal{S}_i^+(\sigma_i)$, we define $\mathcal{P}_{\sigma_i,S}^+$ to be a fixed set of probability measures such that $\text{supp}(\mathcal{P}_{\sigma_i,S}^+) = S$ and $\sigma_i \in \mathfrak{d}_i(\mathcal{P}_{\sigma_i,S}^+)$; likewise, for each $\sigma_i \in \Sigma_i$ and $S \in \mathcal{S}_i^-(\sigma_i)$, define $\mathcal{P}_{\sigma_i,S}^-$ to be a fixed set of probability measures witnessing the fact that $S \in \mathcal{S}_i^-(\sigma_i)$. Then, given $F \in \Omega^\varphi$ with $\text{play}_i(\sigma_i) \in F$, define $\mathcal{PR}_i^\varphi(F)$ as follows: for each $\mu \in \mathcal{P}_{\sigma_i,S}^+ \cup \mathcal{P}_{\sigma_i,S}^-$, let $\tilde{\mu}$ be the unique probability measure on $\text{Bel}_i(F)$ such that, for all $\sigma_{-i} \in \Sigma_{-i}$,

$$\tilde{\mu}(\text{Bel}_i(F; \sigma_{-i})) = \mu(\sigma_{-i}),$$

and which is uniform on each set $\text{Bel}_i(F; \sigma_{-i})$; then set

$$\mathcal{PR}_i^\varphi(F) := \begin{cases} \{\tilde{\mu} : \mu \in \mathcal{P}_{\sigma_i,S}^+\} & \text{if } \text{RAT}_i \in F \\ \{\tilde{\mu} : \mu \in \mathcal{P}_{\sigma_i,S}^-\} & \text{if } \text{RAT}_i \notin F. \blacksquare \end{cases}$$

4.2 Multiple decision rules

In general, we may wish to reason about players who are uncertain about which decision rules their opponents are using. For example, player i might believe that if player j is maximizing her expected utility, she will play σ_j , but if she plays σ'_j , then she might be minimizing the worst-case outcome. One way to try to model such uncertainty is to expand the logic so that there is a set $\mathfrak{D}_i$ of decision rules associated with each player i . Consider the collection of primitive propositions

$$\Phi_\Gamma^\mathfrak{D} := \Phi_\Gamma \cup \{\text{RAT}_i^{\mathfrak{d}_i} : i \in I, \mathfrak{d}_i \in \mathfrak{D}_i\}.$$

Interpret $\text{RAT}_i^{\mathfrak{d}_i}$ as before. Then $\mathcal{L}_B(\Phi_\Gamma^\mathfrak{D})$ is a language for reasoning about the strategies and beliefs of the players $i \in I$ as well as their adherence to the various decision rules $\mathfrak{d}_i \in \mathfrak{D}_i$.

To obtain a complete axiomatization of this language, it is not sufficient to simply let **G4–G6** range over all decision rules $\mathfrak{d}_i \in \mathfrak{D}_i$ for each player i ; in general, this system is sound but not complete. Roughly speaking, this is because it is possible for both $\text{play}_i(\sigma_i) \wedge \text{RAT}_i^{\mathfrak{d}_i}$ and $\text{play}_i(\sigma_i) \wedge \text{RAT}_i^{\mathfrak{d}'_i}$ to be consistent with $\delta_{i,S}$ for some $S \subseteq \Sigma_{-i}$, yet no measure μ with $\text{supp}(\mu) = S$ is such that $\sigma_i \in \mathfrak{d}_i(\mu) \cap \mathfrak{d}'_i(\mu)$. In this case, the formula

$$\text{play}_i(\sigma_i) \wedge \text{RAT}_i^{\mathfrak{d}_i} \wedge \text{RAT}_i^{\mathfrak{d}'_i} \wedge \delta_{i,S}$$

is not satisfiable, but there is no way to prove its negation from the axioms. However, provided that each set $\mathfrak{D}_i$ is finite, we can deal with this problem by replacing **G5** and **G6** with the following collection of axioms for each player i , each strategy $\sigma_i \in \Sigma_i$, and every subset $D \subseteq \mathfrak{D}_i$:

$$\left(\text{play}_i(\sigma_i) \wedge \bigwedge_{\mathfrak{d}_i \in D} \text{RAT}_i^{\mathfrak{d}_i} \wedge \bigwedge_{\mathfrak{d}_i \notin D} \neg \text{RAT}_i^{\mathfrak{d}_i} \right) \rightarrow \bigvee_{S \in \mathcal{S}_i^D(\sigma_i)} \delta_{i,S},$$

where $\mathcal{S}_i^D(\sigma_i)$ is the collection of all $S \subseteq \Sigma_{-i}$ such that there exists a probability measure μ on S such that $\text{supp}(\mu) = S$ and for every $\mathfrak{d}_i \in D$, $\sigma_i \in \mathfrak{d}_i(\mu)$, and for every $\mathfrak{d}_i \notin D$, $\sigma_i \notin \mathfrak{d}_i(\mu)$. **G5** and **G6** are special cases occurring when $|\mathfrak{D}_i| = 1$, corresponding to $D = \mathfrak{D}_i$ and $D = \emptyset$, respectively.

These modified axioms give a sound and complete axiomatization of $\mathcal{L}_B(\Phi_\Gamma^\mathfrak{D})$ with respect to the class of all finite Γ -structures. Nevertheless, there is something unsatisfying about using this language to model players' uncertainty about decision rules: the propositions $RAT_i^{\mathfrak{d}_i}$ say that player i is playing a $\mathfrak{d}_i$ -best-response, but *not* that player i is actually *using* the rule $\mathfrak{d}_i$ to decide her strategy. To see the difference, consider a player i who is trying to maximize her expected utility (i.e. using $\mathfrak{d}_i^{eu}$), and happens to also play a maximin strategy; contrast this with a scenario in which she is actively seeking to maximize the worst-case outcome (i.e. using $\mathfrak{d}_i^m$), and in so doing happens to play a strategy that maximizes her expected utility. Although player i is following different decision rules in these two cases, the language $\mathcal{L}_B(\Phi_\Gamma^\mathfrak{D})$ cannot express this difference; the formula $RAT_i^{\mathfrak{d}_i^{eu}} \wedge RAT_i^{\mathfrak{d}_i^m}$ holds either way. For instance, in the example presented in Section 2.4, it is not hard to check that

$$\omega_0 \models play_1(\text{Bach}) \rightarrow (RAT_1^{\mathfrak{d}_1^{eu}} \wedge RAT_1^{\mathfrak{d}_1^m}).$$

In sum, the propositions $RAT_i^{\mathfrak{d}_i}$ do not say anything about *how* player i is making her decision, but simply record whether or not the decision she does make is compatible with the rule $\mathfrak{d}_i$. What we want is a different kind of proposition, say $rule_i(\mathfrak{d}_i)$, that says that player i really is *using* the rule $\mathfrak{d}_i$ in deciding her strategy.

Decision rules interpreted in this sense are particularly relevant in a dynamic setting. When an opponent does something unexpected and seemingly irrational, there is the question of how to update your beliefs. One option is to abandon the belief that your opponent is rational, but this is unsatisfying both conceptually and methodologically. An alternative response is to update your beliefs about your opponent's beliefs: what they did actually was rational with respect to their beliefs, you had just misjudged what those beliefs were (see, e.g., (Battigalli & Siniscalchi 2002)). But in some cases, this too is unsatisfying or simply unavailable. However, when decision rules are formally present as objects of belief, a third option becomes available: abandon the belief that your opponent is $\mathfrak{d}_i$ -rational, but not that they are behaving rationally with respect to *some* decision rule. Though an analysis of decision rules in extensive-form games is beyond the scope of this paper, the groundwork for such a study can be laid by formalizing them in a static context.

Fix a game Γ and a profile $\mathfrak{D} = (\mathfrak{D}_i)_{i \in I}$ of sets of decision rules for each player $i \in I$. Let

$$\Phi_{\Gamma, \mathfrak{D}} := \Phi_\Gamma \cup \{rule_i(\mathfrak{d}_i) : i \in I, \mathfrak{d}_i \in \mathfrak{D}_i\}.$$

In order to interpret the primitive propositions $rule_i(\mathfrak{d}_i)$, we must extend the semantic model so that it associates with each world ω the decision rule that each player i is using at that world; furthermore, we must constrain the strategies used at each world so that they are compatible with the corresponding decision rules. Formally, a Γ -**structure with decision functions** is a tuple $M = (\Omega, (Pr_i)_{i \in I}, s, r)$ satisfying (C1) through (C4) as well as the following additional conditions:

$$(C5) \ r : \Omega \rightarrow \mathfrak{D} \text{ satisfies } Pr_i(\omega)(\{\omega' \in \Omega : r_i(\omega') = r_i(\omega)\}) = 1;$$

$$(C6) \ s_i(\omega) \in r_i(\omega)(\mu_{i, \omega}).$$

Condition (C5) says that the *decision function* r assigns to each world ω a profile of decision rules $r(\omega)$ —intuitively, $r_i(\omega) \in \mathfrak{D}_i$ is the rule that player i is using at ω —and moreover, each player is sure of her own decision rule. Condition (C6) requires that, at each world ω , the strategy $s_i(\omega)$ is an $r_i(\omega)$ -best response for player i ; in other words, player i really is following the decision rule $r_i(\omega)$ at ω .

The language $\mathcal{L}_B(\Phi_{\Gamma, \mathfrak{D}})$ can be interpreted in a Γ -structure with decision functions M as before, with the additional clause

$$\llbracket rule_i(\mathfrak{d}_i) \rrbracket_M := \{\omega \in \Omega : r_i(\omega) = \mathfrak{d}_i\}.$$

The resulting logic can be axiomatized using essentially the same technique as in Section 3. Consider the following axiom schemes:

- P1. $\bigvee_{\sigma_i \in \Sigma_i} play_i(\sigma_i)$
- P2. $\neg(play_i(\sigma_i) \wedge play_i(\sigma'_i))$, for $\sigma_i \neq \sigma'_i$
- P3. $play_i(\sigma_i) \leftrightarrow B_i play_i(\sigma_i)$
- P4. $\bigvee_{\mathfrak{d}_i \in \mathfrak{D}_i} rule_i(\mathfrak{d}_i)$
- P5. $\neg(rule_i(\mathfrak{d}_i) \wedge rule_i(\mathfrak{d}'_i))$, for $\mathfrak{d}_i \neq \mathfrak{d}'_i$
- P6. $rule_i(\mathfrak{d}_i) \leftrightarrow B_i rule_i(\mathfrak{d}_i)$
- P7. $(play_i(\sigma_i) \wedge rule_i(\mathfrak{d}_i)) \rightarrow \bigvee_{S \in \mathcal{S}_i^+(\sigma_i)} \delta_{i, S}.$

Here, as before, $\mathcal{S}_i^+(\sigma_i)$ denotes the collection of all $S \subseteq \Sigma_{-i}$ such that there exists a probability measure μ with $supp(\mu) = S$ and such that $\sigma_i \in \mathfrak{d}_i(\mu)$. Note that there is no need for a symmetric axiom involving $\mathcal{S}_i^-(\sigma_i)$ for this logic, because the formula $\neg rule_i(\mathfrak{d}_i)$, unlike $\neg RAT_i$, does not say that σ_i is *incompatible* with player i 's beliefs and the decision rule $\mathfrak{d}_i$; it simply says that player i did not *use* the rule $\mathfrak{d}_i$ to help choose her strategy σ_i (though she may, coincidentally, have beliefs with respect to which σ_i is a $\mathfrak{d}_i$ -best response).

Let $GL_{\Gamma, \mathfrak{D}}$ be the axiom system that results from adding P1–P7 to the KD45 axioms and rules of inference of belief logic. Then we have the following result, the proof of which proceeds analogously to that given in Section 3.2.

Theorem 4.2: $GL_{\Gamma, \mathfrak{D}}$ is a sound and complete axiomatization of the language $\mathcal{L}_B(\Phi_{\Gamma, \mathfrak{D}})$ with respect to the class of all finite Γ -structures with decision functions.

5 The role of language

In this section we focus on the profile of decision rules $\mathfrak{d} = \mathfrak{d}^{eu}$, with respect to which each player is rational precisely if they are playing a strategy that maximizes their expected utility. As noted, it is somewhat surprising that G5 and G6 are sufficient to capture this notion of rationality. Whether or not a player is maximizing their expected utility depends on their *quantitative* beliefs; however, while G5 and G6 specify the possible supports for player i 's beliefs,

they say nothing about the actual weights placed on the individual outcomes. Nor could they—the language $\mathcal{L}_B(\Phi_\Gamma^\partial)$ cannot express anything beyond such qualitative properties of the measures $Pr_i(\omega)$.

Expressivity, however, is a double-edged sword: when working with a less expressive language, though we are more limited in the possible axioms we have available, there are also fewer validities to worry about proving. This, in essence, is why GL_Γ^∂ can be a complete axiomatization: the properties of rationality it fails to encode are precisely those properties that are not expressible in the language at all.

A richer language—in particular, one with a finer-grained representation of belief—may not be axiomatizable at all, or at least not using the techniques in this paper. Consider, for example, a language with belief modalities B_i^α for each $\alpha \in [0, 1]$, where $B_i^\alpha \varphi$ is interpreted as saying that player i assigns probability α to φ . In this case, GL_Γ^∂ (replacing B_i by B_i^1) is sound but certainly not complete. It can easily happen, for example, that in the game Γ it is only rational for player i to play σ_i if she assigns probability $\frac{1}{2}$ to σ_{-i} ; however, the corresponding validity

$$(play_i(\sigma_i) \wedge RAT_i) \rightarrow B_i^{\frac{1}{2}} play_{-i}(\sigma_{-i})$$

is clearly not a theorem of GL_Γ^∂ . Moreover, extending GL_Γ^∂ to this richer language runs into difficulties. The axiom schemes **G5** and **G6** essentially work by insisting that the players' beliefs be compatible with rationality or its negation, respectively. In the language $\mathcal{L}_B(\Phi_\Gamma^\partial)$, this amounts to specifying the possible supports for the players' beliefs, which can be written using finite formulas since each Σ_{-i} is finite and therefore has only finitely-many subsets. By contrast, in the language with belief modalities B_i^α for every $\alpha \in [0, 1]$, the “formula” that says that player i 's beliefs are compatible with rationality may be infinitely long.

A still richer language, however, can circumvent these difficulties entirely. Fix a game Γ and consider the language of *linear likelihood inequalities* defined by the grammar

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid a_1 \ell_i(\varphi_1) + \dots + a_k \ell_i(\varphi_k) \geq b,$$

where $p \in \Phi_\Gamma$, $i \in I$, $k \in \mathbb{N}$, and $a_1, \dots, a_k, b \in \mathbb{R}$. The *likelihood terms* $\ell_i(\varphi)$ are meant to be read as “the probability of φ according to player i ”, and a likelihood formula $a_1 \ell_i(\varphi_1) + \dots + a_k \ell_i(\varphi_k) \geq b$ should be thought of as asserting the corresponding inequality. More precisely, we interpret such formulas in a Γ -structure M as follows:

$$\begin{aligned} & \llbracket a_1 \ell_i(\varphi_1) + \dots + a_k \ell_i(\varphi_k) \geq b \rrbracket_M \\ & := \{ \omega \in \Omega : \sum_{j=1}^k a_j Pr_i(\omega)(\llbracket \varphi_j \rrbracket_M) \geq b \}. \end{aligned}$$

For example, the formula $\ell_i(\varphi) \geq 1$ says that player i assigns probability at least (and therefore exactly) 1 to φ , while the formula $(\ell_i(\varphi) \geq \frac{1}{2}) \wedge (\ell_i(\neg\varphi) \geq \frac{1}{2})$ says that player i assigns probability $\frac{1}{2}$ to φ . See (Halpern 2003) for a thorough discussion of this and related logics; a sound and complete axiomatization is given in (Fagin, Halpern, & Megiddo 1990).

In this language, rationality in the sense of expected utility maximization can be *defined*, thus obviating the need for a

separate axiomatization. Indeed, if we let $\sigma_i \succeq \sigma'_i$ be an abbreviation for the formula

$$\sum_{\sigma_{-i} \in \Sigma_{-i}} u_i(\sigma_i, \sigma_{-i}) \ell_i(play_{-i}(\sigma_{-i})) - \sum_{\sigma_{-i} \in \Sigma_{-i}} u_i(\sigma'_i, \sigma_{-i}) \ell_i(play_{-i}(\sigma_{-i})) \geq 0,$$

which says that the expected value according to player i of playing σ_i is no less than the expected value of playing σ'_i , then it is easy to see that

$$RAT_i^{\partial^{eu}} \leftrightarrow \bigvee_{\sigma_i \in \Sigma_i} \left(play_i(\sigma_i) \wedge \bigwedge_{\sigma'_i \in \Sigma_i} \sigma_i \succeq \sigma'_i \right)$$

is valid. Since rationality is expressible in this language, it is axiomatized as well.

6 Further work

Equilibrium analysis in game theory is predicated on the notion of rationality: solution concepts like Nash equilibrium are formulated in terms of best responding to beliefs. As such, one natural application of a logic of rationality is to the analysis of equilibria.

The language $\mathcal{L}_B(\Phi_\Gamma)$ has formulas that represent pure strategy choices, but not mixed strategies. In the context of Nash equilibrium, this difference turns out to be (at least formally) innocuous: one can view a mixed strategy $\ell_i \in \Delta(\Sigma_i)$ either as a conscious randomization on the part of player i , or as the common conjecture of the players $j \neq i$ about what pure strategy i will choose—either way, the set of mixed strategy Nash equilibria stays the same. However, this insensitivity is, in part, dependent on that fact that rationality in the sense of expected utility maximization plays well with mixing: $\ell_i \in \Delta(\Sigma_i)$ maximizes player i 's expected utility (with respect to some fixed beliefs) if and only if every pure strategy σ_i in the support of ℓ_i maximizes expected utility. But this correspondence breaks down when “expected utility maximization” is replaced with the generalized notion of rationality presented in Section 2.3: in the context of an arbitrary decision rule $\mathfrak{d}_i : \Delta(\Sigma_{-i}) \rightarrow 2^{\Sigma_i}$, there is no principled way to extend the notion of “best response” from Σ_i to $\Delta(\Sigma_i)$. This suggests that further research into the interaction between pure and mixed strategies under general decision rules may be fruitful.

One might also wonder whether there are useful logical characterizations of when a given strategy constitutes a Nash equilibrium. Similar to the above, a related line of inquiry arises by considering generalized decision rules and the correspondingly generalized notions of equilibria they induce. In principle, it makes sense to consider “ $\mathfrak{d}$ -Nash equilibria”, defined by replacing $\mathfrak{d}^{eu}$ with an arbitrary profile of decision rules $\mathfrak{d}$ in the definition of Nash equilibrium. It is certainly too much to hope that Nash's famous existence theorem apply in full force to this wider concept; however, properties of $\mathfrak{d}$ that suffice to guarantee the existence of equilibria are of interest, and potentially admit a logical characterization. Such questions are the subject of ongoing research.

Although in this paper we focused on rationality, our approach to axiomatizing rationality applies to a wider class of properties. Define an *i -subjective property* to be any property of i that i is sure of; that is, a formula φ such that

$\varphi \leftrightarrow B_i \varphi$ is valid. Rationality is i -subjective, as is whether i is jealous (e.g., believes somebody has some property that she doesn't have, but wishes she did), whether i is cautious (e.g., assigns positive probability to all her opponents' actions), etc. We believe that the techniques developed in this paper could be used to provide a sound and complete axiomatization of any i -subjective property whose truth is determined by i 's beliefs and what i plays. We hope to return to this topic in future work.

7 Acknowledgements

Bjorndahl is supported in part by NSF grants IIS-0812045, CCF-1214844, DMS-0852811, and DMS-1161175, and ARO grant W911NF-09-1-0281. Halpern is supported in part by NSF grants IIS-0812045, IIS-0911036, and CCF-1214844, by MURI grant FA9550-12-1-0040, and by ARO grant W911NF-09-1-0281. Pass is supported in part by an Alfred P. Sloan Fellowship, a Microsoft Research Faculty Fellowship, NSF Awards CNS-1217821 and CCF-1214844, NSF CAREER Award CCF-0746990, AFOSR YIP Award FA9550-10-1-0093, and DARPA and AFRL under contract FA8750-11-2-0211. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

References

- Battigalli, P., and Siniscalchi, M. 2002. Strong beliefs and forward-induction reasoning. *Journal of Economic Theory* 106:356–391.
- Bjorndahl, A.; Halpern, J. Y.; and Pass, R. 2011. Reasoning about justified belief. In *Theoretical Aspects of Rationality and Knowledge: Proc. Thirteenth Conference (TARK 2011)*, 221–227.
- Brandenburger, A., and Dekel, E. 1987. Rationalizability and correlated equilibria. *Econometrica* 55:1391–1402.
- Fagin, R.; Halpern, J. Y.; Moses, Y.; and Vardi, M. Y. 1995. *Reasoning About Knowledge*. Cambridge, Mass.: MIT Press. A slightly revised paperback version was published in 2003.
- Fagin, R.; Halpern, J. Y.; and Megiddo, N. 1990. A logic for reasoning about probabilities. *Information and Computation* 87(1/2):78–128.
- Halpern, J. Y., and Moses, Y. 2007. Characterizing solution concepts in games using knowledge-based programs. In *Proc. Twentieth International Joint Conference on Artificial Intelligence (IJCAI '07)*, 1300–1307.
- Halpern, J. Y. 2003. *Reasoning About Uncertainty*. Cambridge, Mass.: MIT Press.
- Hayashi, T. 2008. Regret aversion and opportunity dependence. *Journal of Economic Theory* 139(1):242–268.
- Osborne, M. J., and Rubinstein, A. 1994. *A Course in Game Theory*. Cambridge, Mass.: MIT Press.
- Tan, T., and Werlang, S. 1988. The Bayesian foundation of solution concepts of games. *Journal of Economic Theory* 45(45):370–391.