Reasoning About Justified Belief[*]

Adam Bjorndahl
Cornell University
Dept. Mathematics
Ithaca, NY 14853, USA
abjorndahl@math.cor-
nell.edu

Joseph Y. Halpern
Cornell University
Dept. Computer Science
Ithaca, NY 14853, USA
halpern@cs.cornell.edu

Rafael Pass
Cornell University
Dept. Computer Science
Ithaca, NY 14853, USA
rafael@cs.cornell.edu

ABSTRACT

Halpern and Pass [8] introduce a logic of *justified belief* and go on to prove that *strong rationalizability* is characterized in this logic in terms of *common justified belief of rationality* (CJBR). Their paper provides semantics for this logic but no axiomatization. We correct this deficiency by reformulating the definition of justified belief and providing a complete axiomatization of this new system. We then prove a result analogous to the characterization of strong rationalizability in terms of CJBR, and analyze the additional assumptions needed to do so.

Categories and Subject Descriptors

F.4.1 [**Mathematical Logic and Formal Languages**]: Mathematical Logic—*modal logic*; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*multiagent systems*; J.4 [**Social and Behavioral Sciences**]: Economics

General Terms

Economics, Theory

Keywords

Epistemic logic, rationality, justified belief

1. INTRODUCTION

One of the best known solution concepts in game theory is *rationalizability* [9]. Roughly speaking, a strategy σ for player i is rationalizable if σ is a best response to some belief of player i about the strategies of the other players, under the assumption that these strategies too are rationalizable (so are themselves best responses to players' beliefs, and so on). As shown by Tan and Werlang [11] and Brandenburger and Dekel [3], a strategy is rationalizable if and only if it can be played at a state where rationality is common knowledge.

However, it is known that there is a sense in which rationalizability is too permissive. For example, in the well-known *centipede game* [10], viewed as a normal-form game, every strategy is rationalizable, despite a backward-induction argument that yields a unique course of action for the

starting player: quitting immediately.[1]

The culprit here seems to be that the two players may have "incompatible" beliefs. If the second player (Bob) believes the first player (Alice) will quit immediately (i.e. if Bob assigns probability 1 to that event), then it is easily seen that Bob can rationalize any strategy, since what he does has no influence on the outcome. But if Bob can rationalize any strategy, then so can Alice; for example, if Bob chooses to quit at round 4 (believing that Alice will quit at round 1), then Alice can rationalize quitting at round 3. A key observation in this example is that while Alice assigns probability 1 to the event "Alice quits at round 3", Bob assigns this same event probability 0.

This suggests that a strengthening of the notion of rationalizability might be fruitful and, moreover, that the missing component is some sort of "compatibility" condition on the players' beliefs. This strengthening is realized in the definition of *strong rationalizability* given by Halpern and Pass [8] (HP from now on). Roughly speaking, a strategy is strongly rationalizable if it is rationalizable using beliefs that are compatible, in the sense that if one player assigns positive probability to a strategy profile, then all players assign positive probability to it. HP show that in the centipede game, the only strategy profiles that are strongly rationalizable are ones where Alice quits right away.

HP analyze strong rationalizability from a modal perspective, using a logic that includes distinct modal operators for *belief* and *justified belief*. Among other things, they show that common justified belief of rationality (CJBR) characterizes strongly rationalizable strategies. Roughly speaking, according to the HP definition, player i has justified belief in φ at a state ω, denoted $\omega \models B_i^* \varphi$, if (a) the player believes φ (that is, he assigns φ probability 1), and (b) his probability distribution gives ω positive probability. The problem with this definition is that at a state ω that is not given positive probability, the player cannot have justified belief of anything; even $B_i^* true$ does not hold. To deal with this, HP take $B_i^* \varphi$ to be true at ω if either the player believes φ and gives ω positive probability, or φ is valid.

While this approach does ensure that $B_i^* true$ holds at every state, and does suffice to allow HP to characterize strong

---

[1]Recall that in the centipede game two players take turns moving; at each move before the end, they can either quit the game or continue (the player who moves at the last step can only quit). For all steps t before the end, the player that moves at t prefers the outcome of stopping at round t to the outcome of stopping at round $t+1$. However, for all t, stopping in round $t+2$ leads to a better outcome for both players than stopping in round t.

rationalizability in terms of CJBR, it seems somewhat *ad hoc*. Moreover, the HP definition proves difficult to generalize or adapt to differing intuitions. For example, it seems reasonable to require that a player has justified belief of his own beliefs, but there is no natural way to incorporate this requirement into the definition short of simply imposing it as a third disjunct.

We provide here a reformulation of the logic of justified belief that, while remaining true to the original intent, is a more natural object of study. For example, it allows us to naturally capture the requirement that players have justified belief of their own beliefs. Further evidence of the naturalness of our approach is provided by a straightforward sound and complete axiomatization of the logic, a feature lacking in [8].

Our reformulation is based on the idea of a logic that includes, in addition to an epistemic belief modality B_i, an *alethic* modality $\square_i$ (that lets us talk about logical necessity and possibility); B_i^* is defined as the conjunction of B_i and $\square_i$. This idea is somewhat reminiscent of an approach taken by Artemov and Nogin [2], who have also considered adding justification to epistemic logic. They do so by use of "justification terms" of the form $t : \varphi$. The formula $t : \varphi$ says that t is an explicit witness (perhaps a proof) that φ is true. The formula also has an epistemic component. The implicit assumption is that an agent who has a proof understands that he has a proof, so that the implication $(t : \varphi) \Rightarrow B\varphi$ holds. No such implication holds for our $\square_i$ modality. $B_i^* \varphi$ is perhaps best thought of as "exists t such that $t :_i \varphi$ holds" (where $t :_i \varphi$ says that "agent i has a justification t for φ"). (Artemov [1] and Fitting [5] have considered translating formulas with explicit justifications into a simple modal logic with a $\square$ operator by replacing each term of the form $t : \varphi$ by $\square \varphi$.) It would be interesting to see if there are axioms on justification that would allow us to reproduce the properties of B_i^* using this approach, although doing so is beyond the scope of this paper.

Our new approach does have a downside: it is no longer the case in general that CJBR characterizes strong rationalizability in all structures. But this failing can be mitigated: we can identify exactly which additional properties are needed in a structure to recover the characterization. This leads to a deeper understanding of both justified belief and strong rationalizability.

The rest of the paper is organized as follows. In the next section we review the semantics of the HP notion of justified belief and discuss the shortcomings of this approach. In Section 3 we motivate, define, and completely axiomatize a new formulation of justified belief; we then analyze some of its fundamental properties. The discussion in Sections 2 and 3 is carried out in a general logical setting. In Section 4, we specialize the setting to games, expanding the language to include formulas for strategies and rationality; we then provide a sound and complete axiomatization. Thus, perhaps surprisingly, we can reason about rationality in a qualitative language that can talk only about beliefs, rather than probability and utility. In Section 5, we characterize the key properties of justified belief needed to prove the characterization of strong rationalizability in terms of CJBR. This turns out to require extra "richness" requirements; Section 6 is therefore devoted to exploring some of the consequences of these additional requirements, how they might be viewed in a wider context, and directions for future research.

2. THE HALPERN-PASS DEFINITION OF JUSTIFIED BELIEF

We begin by establishing a basic logical setting in which the justified belief operator may be defined.

Let $\mathcal{L}_n^B(\Phi)$ denote the language that has primitive propositions in Φ, and is closed under the standard Boolean connectives as well as the modal operators B_i ("player i believes that"), for $1 \le i \le n$. (As usual, we omit Φ and write just $\mathcal{L}_n^B$ when Φ is not relevant.) We use Kripke-style semantics, where associated to each state ω and each player i is a probability measure on the state space, thought of as representing player i's beliefs at ω. Formally, a *probability frame* is a tuple $(\Omega, \mathcal{PR}_1, \ldots, \mathcal{PR}_n)$ satisfying the following conditions:

(P1) Ω is a nonempty topological space;

(P2) each $\mathcal{PR}_i$ assigns to each $\omega \in \Omega$ a probability measure $\mathcal{PR}_i(\omega)$ on Ω;

(P3) $\omega' \in \mathcal{PR}_i[\omega] \Rightarrow \mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega)$, where $\mathcal{PR}_i[\omega]$ abbreviates $Supp(\mathcal{PR}_i(\omega))$, the support of the probability measure.

Condition (P3) ensures that each player is sure of his own beliefs. The topological structure on Ω is necessary to make sense of the probability measures, which are implicitly taken to be defined on the Borel subsets of Ω. For simplicity, in this abstract, we will restrict our attention to finite state spaces with the discrete topology, in which case all subsets of Ω are measurable, so we can suppress mention of the topological structure altogether. In the full paper, we extend our results to infinite state spaces.

A *probability structure M* is a probability frame together with a *valuation function* $[\![\cdot]\!]_M : \Phi \to 2^\Omega$. This valuation is extended to all formulas recursively via:

$$
\begin{array}{lll}
[\![\varphi \wedge \psi]\!]_M & =_{\text{def}} & [\![\varphi]\!]_M \cap [\![\psi]\!]_M \\
[\![\neg \varphi]\!]_M & =_{\text{def}} & \Omega - [\![\varphi]\!]_M \\
[\![B_i \varphi]\!]_M & =_{\text{def}} & \{ \omega \in \Omega : \mathcal{PR}_i[\omega] \subseteq [\![\varphi]\!]_M \}.
\end{array}
$$

Thus, the Boolean connectives are interpreted classically, while the formula $B_i \varphi$ holds at all states ω such that $\mathcal{PR}_i(\omega)$ assigns probability 1 to φ. As is standard, we often write $(M, \omega) \models \varphi$ or just $\omega \models \varphi$ for $\omega \in [\![\varphi]\!]_M$; similarly, we write $M \models \varphi$ for $[\![\varphi]\!]_M = \Omega$; and we say that φ is *valid*, and write $\models \varphi$, if $M \models \varphi$ for all probability structures M. When $(M, \omega) \not\models \varphi$, we say that M *refutes φ at ω*, or just that ω refutes φ.

The goal now is to introduce a second unary modal operator for each player, B_i^*, to be interpreted in some sense as "justified belief". As a first attempt at providing semantics for this operator, consider:

$$ \omega \models B_i^* \varphi \Leftrightarrow \omega \models B_i \varphi \text{ and } \omega \in \mathcal{PR}_i[\omega]. $$

This is meant to capture the intuition that a justified belief should never rule out something which might, in fact, obtain; to put the same point evidentially, one cannot justifiably discount anything that one has not observed evidence against. In particular, then, a justified belief must include the actual state in its support.

These semantics for B_i^*, however, yield a non-normal operator, since B_i^* need not hold even of valid formulas: at any state ω with $\omega \notin \mathcal{PR}_i[\omega]$, $B_i^* \varphi$ fails for all φ. This motivates changing the semantics for B_i^* to

$$ \omega \models B_i^* \varphi \; \Leftrightarrow \; \begin{array}{ll} \text{(a)} & \omega \models B_i \varphi \text{ and } \omega \in \mathcal{PR}_i[\omega], \text{ or} \\ \text{(b)} & \varphi \text{ is valid,} \end{array} $$

which is precisely the HP definition.

This solves one problem but creates another: axiomatizing B_i^* now seems to require an ability to express "is valid" in the object language, since the formula

$$B_i^* \varphi \Rightarrow (B_i \psi \Rightarrow \psi) \tag{1}$$

is valid for all and only refutable φ. Thus, it seems that in order to axiomatize B_i^* we need to reason about satisfiability and validity. While this can be done [6], it seems not to get at the essence of justified belief. It can also be argued that a player ought to have justified belief in his own beliefs at all states, not just those which lie in the support of their own probability measure. This can be captured by adding a (rather ugly!) third disjunct to an already *ad hoc* definition, but again, this does not seem to be a natural way to go. This motivates the main goal of this paper: to present an alternative approach to defining the B_i^* operators that resolves all of the issues raised above.

3. A NEW APPROACH TO JUSTIFIED BELIEF

Returning to the language $\mathcal{L}_n^B$, we begin, not with B_i^*, but by introducing unary modal operators $\square_i$ for each player i. The idea underlying this move stems from the intuition for justified belief given above: that it should never rule out a possibility that *might* obtain. The operators $\square_i$ are intended to introduce the dimension of alethic modality that the word "might" carries in this intuition. As such, we *define* the symbols B_i^* into our language via

$$B_i^* \varphi =_{\text{def}} B_i \varphi \wedge \square_i \varphi,$$

and henceforth take this as our definition of justified belief. We might read $\square_i \varphi$ as "necessarily φ" and, dually, $\diamondsuit_i \varphi =_{\text{def}} \neg \square_i \neg \varphi$ as "possibly φ" or "it might be the case that φ", though we shall see that these readings still require refinement. Loosely speaking, then, we can read the above as: "player i has justified belief in φ just in case player i believes φ and, moreover, it is not the case that $\neg \varphi$ might obtain". Call this new language $\mathcal{L}_n^{B,\square}$.

We no longer need to give semantics to B_i^* directly, since it will inherit its semantics from B_i and $\square_i$. We let $R_i[\omega]$ denote $\{\omega' \in \Omega : \omega R_i \omega'\}$, and define a B^*-*frame* to be a tuple $(\Omega, \mathcal{PR}_1, \ldots, \mathcal{PR}_n, R_1, \ldots, R_n)$ where $(\Omega, \mathcal{PR}_1, \ldots, \mathcal{PR}_n)$ is a probability frame, and the following conditions hold:

(F1) each R_i is a reflexive, transitive relation on Ω;

(F2) $\omega' \in \mathcal{PR}_i[\omega] \Rightarrow R_i[\omega'] \subseteq \mathcal{PR}_i[\omega]$;

(F3) $\omega R_i \omega' \Rightarrow \mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega)$.

A B^*-frame $(\Omega, \overrightarrow{\mathcal{PR}}, \overrightarrow{R})$ is said to be *based on* the probability frame $(\Omega, \overrightarrow{\mathcal{PR}})$. Each R_i is called an *accessibility relation*; we think of it as telling us which states are "possible" or "imaginable" from which other states, and condition (F1) is standard in this respect. Condition (F2) expresses a one-directional transparency between the alethic modality, "it might be the case that" and its epistemic counterpart, "it is considered possible that"; namely, no player considers it possible that something *might* be true without also considering it possible that it *is* true. This may be considered a "dictum of responsible imagining"[2]: if you consider something impossible, then you cannot imagine a world where it

[2]We thank Christina Bjorndahl for suggesting this phrasing.

is possible. Finally, condition (F3) expresses an additional restriction on alethic possibility: that it is possibility *conditioned on* the player's actual beliefs. That is, any state that is possible (from a given state) is a state in which the player's beliefs are the same (as in the given state). As we shall see, it is (F3) that guarantees that an agent has justified beliefs about his own beliefs.

We define a B^*-*structure* to be a B^*-frame together with a valuation function as described above, extended by the additional rule

$$\llbracket \square_i \varphi \rrbracket_M =_{\text{def}} \{\omega \in \Omega : R_i[\omega] \subseteq \llbracket \varphi \rrbracket_M\}.$$

Thus, $\square_i \varphi$ holds at a world ω if φ holds at all worlds that are R_i-accessible from ω.

The following result shows that if $B_i^* \varphi$ holds according to the HP definition, then it also holds according to our current formulation.

PROPOSITION 1. *Let* $(\Omega, \overrightarrow{\mathcal{PR}}, \overrightarrow{R}, \llbracket \cdot \rrbracket_M)$ *be a* B^*-*structure. For all* $\omega \in \Omega$, *each* $1 \leq i \leq n$, *and any formula* φ, *if either*

(a) $\omega \models B_i \varphi$ *and* $\omega \in \mathcal{PR}_i[\omega]$, *or*

(b) $\llbracket \varphi \rrbracket_M = \Omega$, *or*

(c) φ *has the form* $B_i \psi$ *or* $\neg B_i \psi$ *and* $\omega \models \varphi$,

then $\omega \models B_i^* \varphi$.

PROOF. First suppose that (a) holds. Clearly it suffices to show that $\omega \models \square_i \varphi$. Let $\omega' \in \Omega$ be such that $\omega R_i \omega'$; then condition (F2) ensures that $\omega' \in \mathcal{PR}_i[\omega]$, which implies $\omega' \models \varphi$, thereby establishing $\omega \models \square_i \varphi$. Case (b) is obvious. If (c) holds, observe that by condition (P3) we have $\omega \models B_i \varphi$, and condition (F3) guarantees that $\omega \models \square_i \varphi$. (This is true both if φ has the form $B_i \psi$ and if it has the form $\neg B_i \psi$.) $\square$

According to the HP definition, $\omega \models B_i^* \varphi$ if either (a) $\omega \models B_i \varphi$ and $\omega \in \mathcal{PR}_i[\omega]$ or (b) φ is valid. Parts (a) and (b) of Proposition 1 show that $B_i^* \varphi$ continues to hold in either of these two cases under the new definition. Proposition 1(c) shows that each player also has justified belief in his own beliefs: both $B_i \varphi \Rightarrow B_i^* B_i \varphi$ and $\neg B_i \varphi \Rightarrow B_i^* \neg B_i \varphi$ are valid. The proof of Proposition 1 shows that (F3) is crucial for these properties.

As we now show, the logic has a straightfoward axiomatization with the B_i as KD45 operators, the $\square_i$ as S4 operators, and two interaction axiom schemes which capture conditions (F2) and (F3).

Axiom Schemes:

CPC.   All tautologies of classical logic

K$_B$. $B_i(\varphi \Rightarrow \psi) \Rightarrow (B_i \varphi \Rightarrow B_i \psi)$
D. $B_i \varphi \Rightarrow \neg B_i \neg \varphi$
4$_B$. $B_i \varphi \Rightarrow B_i B_i \varphi$
5. $\neg B_i \varphi \Rightarrow B_i \neg B_i \varphi$

K$_\square$. $\square_i(\varphi \Rightarrow \psi) \Rightarrow (\square_i \varphi \Rightarrow \square_i \psi)$
T. $\square_i \varphi \Rightarrow \varphi$
4$_\square$. $\square_i \varphi \Rightarrow \square_i \square_i \varphi$

I1. $B_i \varphi \Rightarrow B_i \square_i \varphi$
I2. $\diamondsuit_i B_i \varphi \Rightarrow \square_i B_i \varphi$

Rules of Inference:

MP. From $\varphi \Rightarrow \psi$ and φ infer ψ
N$_B$. From φ infer $B_i\varphi$
N$_\square$. From φ infer $\square_i\varphi$

Let AX^{B^*} consist of the axioms and rules of inference above.

THEOREM 1. AX^{B^*} *is a sound and complete axiomatization of the language* $\mathcal{L}_n^{B,\square}$ *with respect to the class of all* B^**-structures.*

Soundness is proved as usual, by induction on the length of the deduction; in particular, condition (F2) guarantees **I1** and (F3) guarantees **I2**, as is easily checked. Completeness can be proved by the canonical model method. The (quite standard) details are left to the full paper.

Since the B_i^* operators are defined into our language, they do not occur in the axiomatization AX^{B^*}. We catalogue some of their properties here.

PROPOSITION 2. *The following formulas are valid:*

(a) $B_i^*(\varphi \Rightarrow \psi) \Rightarrow (B_i^*\varphi \Rightarrow B_i^*\psi)$;

(b) $B_i^*\varphi \Rightarrow \varphi$;

(c) $B_i^*\varphi \Rightarrow B_i^*B_i^*\varphi$;

(d) $B_i\varphi \Rightarrow B_iB_i^*\varphi$.

PROOF. Part (a) is a routine verification. Part (b) follows easily from the axiom scheme **T**. Part (c) follows from the fact that $B_i\varphi$ and $\square_i\varphi$ together imply $B_iB_i\varphi$, $B_i\square_i\varphi$, $\square_iB_i\varphi$, and $\square_i\square_i\varphi$, as witnessed by the axiom schemes **4$_B$**, **I1**, **I2**, and **4$_\square$**, respectively. Part (d) is perhaps best observed by noting that if $\omega \models B_i\varphi$ and $\omega' \in \mathcal{PR}_i[\omega]$, then $\omega' \models B_i\varphi$ and $\omega' \in \mathcal{PR}_i[\omega']$; the result now follows by Proposition 1(a). $\square$

The absence of a theorem corresponding to axiom **5**, negative introspection for B_i^*, is no accident. In fact, even the weaker formula

$$\neg B_i^*\varphi \Rightarrow B_i\neg B_i^*\varphi \qquad (2)$$

is not valid: if $\omega \models B_i\varphi \wedge \neg\square_i\varphi$, then certainly $\omega \models \neg B_i^*\varphi$. However, Proposition 2(d) guarantees that $\omega \models B_iB_i^*\varphi$, which of course implies that $\omega \not\models B_i\neg B_i^*\varphi$. It is worth noting (and easy to check) that justified belief in the HP sense also satisfies Proposition 2, and also fails to satisfy both negative introspection and the weaker formulation given in (2).

Having explored some of the syntactic properties of the B_i^* operators, we turn now to a closer examination of B^*-structures; specifically, we are interested in the role that the relations R_i play in determining the nature of justified belief. For example, observe that the identity relation satisfies conditions (F1) through (F3). If M is a B^*-structure in which R_i is the identity, then $M \models \varphi \Leftrightarrow \square_i\varphi$, and therefore $M \models B_i^*\varphi \Leftrightarrow (B_i\varphi \wedge \varphi)$. Thus, the notion of justified belief we have defined subsumes the notion of true belief. Moreover, as is well known, R_i being the identity is characterized syntactically by the axiom scheme

$$\varphi \Rightarrow \square_i\varphi.$$

If we view the identity relation as the minimal relation satisfying (F1) through (F3), then we are naturally led to the investigation of a "maximal" such relation. This is a notion that will play an important role for us in Section 5.

Given a probability frame $(\Omega, \overrightarrow{\mathcal{PR}})$, define the relations Q_i as follows: for all $\omega, \omega' \in \Omega$,

$$\omega Q_i\omega' \Leftrightarrow \begin{array}{ll} \text{(a)} & \mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega), \text{ and} \\ \text{(b)} & \omega \in \mathcal{PR}_i[\omega] \Rightarrow \omega' \in \mathcal{PR}_i[\omega]. \end{array} \qquad (3)$$

This definition ensures that $\omega Q_i\omega'$ holds whenever it does not violate conditions (F2) or (F3); intuitively, this should make the Q_i "as big as possible". This intuition is borne out in the following proposition.

PROPOSITION 3. *The tuple* $(\Omega, \overrightarrow{\mathcal{PR}}, Q_1, \ldots, Q_n)$ *is a* B^**-frame. In fact, it is the unique* B^**-frame based on the probability frame* $(\Omega, \overrightarrow{\mathcal{PR}})$ *in which each accessibility relation* Q_i *is maximal with respect to inclusion.*

PROOF. We first need to verify conditions (F1) through (F3). It is immediate from the definition that each Q_i is reflexive, and transitivity is likewise straightforward; the remaining conditions are satisfied trivially by definition of the Q_i. Maximality and uniqueness are now evident. $\square$

A B^*-frame $(\Omega, \overrightarrow{\mathcal{PR}}, \overrightarrow{R})$ is called *maximal* if the relations R_i are maximal in the sense of Proposition 3. Like minimality, maximality can be characterized syntactically. Let $\mathrm{AX}_{max}^{B^*}$ be the system AX^{B^*} together with the following two axiom schemes:

M1. $\square_i\varphi \Rightarrow B_i\varphi$
M2. $\varphi \Rightarrow \square_i((B_i\psi \wedge \neg\psi) \Rightarrow \Diamond_i\varphi)$

M1 says that for each state ω, $R_i[\omega] \supseteq \mathcal{PR}_i[\omega]$. **M2** is perhaps best understood as an augmented version of the standard axiom that characterizes symmetric relations R_i: $\varphi \Rightarrow \square_i\Diamond_i\varphi$. The formula $B_i\psi \wedge \neg\psi$ cannot hold at any state ω satisfying $\omega \in \mathcal{PR}_i[\omega]$; thus **M2** says essentially that $\omega R_i\omega'$ implies $\omega' R_i\omega$ whenever $\omega' \notin \mathcal{PR}_i[\omega']$.

THEOREM 2. $\mathrm{AX}_{max}^{B^*}$ *is a sound and complete axiomatization of the language* $\mathcal{L}_n^{B,\square}$ *with respect to the class of all maximal* B^**-structures.*

If M is a maximal B^*-structure, then Theorem 2 implies that $M \models B_i^*\varphi \Leftrightarrow \square_i\varphi$. Loosely speaking, then, while at one extreme our notion of justified belief collapses to (merely) true belief, at the other extreme it is realized as full-fledged necessity.

Up to now, we have not included common (justified) belief in the language, so as to focus on the main issues involved in defining justified belief. Common justified belief is needed for the HP characterization of strong rationalizability, however, so we now add it to the language.

Let $\mathcal{L}_n^{CB^*}$ be the language that results from adding the operators CB (common belief) and CB^* (common justified belief) to $\mathcal{L}_n^{B,\square}$ and closing off under all the operators. To give semantics to these new operators, we make use of the following (standard) abbreviations

$$\begin{array}{ll} EB^1\varphi & =_{\mathrm{def}} \quad B_1\varphi \wedge \cdots \wedge B_n\varphi \\ EB^k\varphi & =_{\mathrm{def}} \quad EB(EB^{k-1}\varphi) \\ (EB^*)^1\varphi & =_{\mathrm{def}} \quad B_1^*\varphi \wedge \cdots \wedge B_n^*\varphi \\ (EB^*)^k\varphi & =_{\mathrm{def}} \quad EB^*((EB^*)^{k-1}\varphi), \end{array}$$

and we extend the valuation as follows:

$$\llbracket CB\varphi \rrbracket_M \quad =_{\text{def}} \quad \bigcap_{k=1}^{\infty} \llbracket EB^k \varphi \rrbracket_M$$

$$\llbracket CB^*\varphi \rrbracket_M \quad =_{\text{def}} \quad \bigcap_{k=1}^{\infty} \llbracket (EB^*)^k \varphi \rrbracket_M.$$

Thus, as usual, common belief of φ means that everyone believes that everyone believes $\dots \varphi$; common justified belief is defined analogously.

The axioms for common belief and common justified belief are just variants of the standard fixed-point axiom and induction rule for common knowledge [4; 7].

Axiom Schemes:

FPA$_{CB}$. $\quad CB\varphi \Rightarrow EB(\varphi \wedge CB\varphi)$
FPA$_{CB^*}$. $\quad CB^*\varphi \Rightarrow EB^*(\varphi \wedge CB^*\varphi)$

Rules of Inference:

IR$_{CB}$. $\quad$ From $\psi \Rightarrow EB(\varphi \wedge \psi)$ infer $\psi \Rightarrow CB\varphi$
IR$_{CB^*}$. $\quad$ From $\psi \Rightarrow EB^*(\varphi \wedge \psi)$ infer $\psi \Rightarrow CB^*\varphi$

Let AX^{CB^*} be the system that results from adding these axioms and rules of inference to AX^{B^*}. Using standard techniques [4; 7], we can prove the following result.

THEOREM 3. AX^{CB^*} *is a sound and complete axiomatization of the language* $\mathcal{L}_n^{CB^*}$ *with respect to the class of all* B^**-structures.*

4. STRUCTURES APPROPRIATE FOR GAMES

We now want to apply justified belief to game theory, with the goal of characterizing strong rationalizability by CJBR, as in [8].

Fix a normal-form n-player game Γ, where $\Sigma_i(\Gamma)$ denotes the strategies of player i in Γ,

$$\Sigma(\Gamma) := \prod_{i=1}^{n} \Sigma_i(\Gamma),$$

and

$$\Sigma_{-i}(\Gamma) := \prod_{j \neq i} \Sigma_j(\Gamma).$$

To reason about players' actions and rationality in Γ, following HP, we take Φ_Γ to consist of the primitive propositions $play_i(\sigma_i)$ for $\sigma_i \in \Sigma_i(\Gamma)$ ("player i is playing strategy σ_i") and RAT_i ("player i is rational"), and consider the language $\mathcal{L}_n^{CB^*}(\Phi_\Gamma)$; to simplify notation, we write $\mathcal{L}_n^{CB^*}(\Gamma)$ rather than $\mathcal{L}_n^{CB^*}(\Phi_\Gamma)$. We make use of the following syntactic abbreviations:

$$RAT \quad =_{\text{def}} \quad RAT_1 \wedge \cdots \wedge RAT_n$$
$$play(\overrightarrow{\sigma}) \quad =_{\text{def}} \quad play_1(\sigma_1) \wedge \cdots \wedge play_n(\sigma_n).$$

A *structure appropriate for* Γ (or a Γ*-structure* for short) is essentially a B^*-structure that interprets these primitive propositions appropriately. Formally, it is a tuple $M = (\Omega, \mathbf{s}, \overrightarrow{\mathcal{PR}}, \overrightarrow{R})$ where $(\Omega, \overrightarrow{\mathcal{PR}}, \overrightarrow{R})$ is a B^*-frame and $\mathbf{s}$ is a *strategy function* that associates to each state $\omega \in \Omega$ a pure strategy profile $\mathbf{s}(\omega) \in \Sigma(\Gamma)$ satisfying

(S1) $\mathcal{PR}_i[\omega] \subseteq \llbracket \mathbf{s}_i(\omega) \rrbracket_M$,

where $\mathbf{s}_i(\omega)$ denotes player i's strategy in the strategy profile $\mathbf{s}(\omega)$ and

$$\llbracket \sigma_i \rrbracket_M := \{\omega : \mathbf{s}_i(\omega) = \sigma_i\}$$

(so $\llbracket \mathbf{s}_i(\omega) \rrbracket_M = \{\omega' : \mathbf{s}_i(\omega') = \mathbf{s}_i(\omega)\}$). Condition (S1) ensures that player i is sure of his own strategy. It is worth noting that all Γ-structures also satisfy

(S2) $R_i[\omega] \subseteq \llbracket \mathbf{s}_i(\omega) \rrbracket_M$,

which shows that the alethic notion of possibility captured by the relation R_i is possibility *conditioned on* player i's actual strategy. This parallels condition (F3), and ensures that each player has justified belief not only of his own beliefs, but also of his own strategy. If we omit the relations R_i, we obtain the HP definition of a *probability structure appropriate for* Γ (or a *probability* Γ*-structure* for short).

The function $\mathbf{s}$ induces a valuation $\llbracket \cdot \rrbracket_M : \Omega \to 2^\Omega$ on primitive propositions as follows:

$$\llbracket play_i(\sigma_i) \rrbracket_M \quad =_{\text{def}} \quad \llbracket \sigma_i \rrbracket_M$$
$$\llbracket RAT_i \rrbracket_M \quad =_{\text{def}} \quad \{\omega \in \Omega : \mathbf{s}_i(\omega) \text{ is a best response given } \mathcal{PR}_i(\omega)\},$$

where the notion of "best response" is determined in Γ according to player i's beliefs on the strategies of other players induced by $\mathcal{PR}_i(\omega)$. More precisely, a probability measure π on Ω induces a probability measure μ on $\Sigma_{-i}(\Gamma)$ via

$$\mu(\sigma_{-i}) = \pi(\llbracket \sigma_{-i} \rrbracket_M),$$

where σ_{-i} and $\llbracket \sigma_{-i} \rrbracket_M$ are defined in the obvious way. The measure μ can then be combined with the utility function given by the game Γ to generate a notion of "best response" via expected utility.

We next provide axioms that characterize the interpretation of $play_i(\sigma_i)$ and RAT_i in Γ-structures. Given $S \subseteq \Sigma_{-i}(\Gamma)$ and $\sigma_{-i} \in \Sigma_{-i}(\Gamma)$, let

$$\chi_S(\sigma_{-i}) =_{\text{def}} \begin{cases} \neg B_i \neg play_{-i}(\sigma_{-i}) & \text{if } \sigma_{-i} \in S \\ B_i \neg play_{-i}(\sigma_{-i}) & \text{if } \sigma_{-i} \notin S. \end{cases}$$

Axiom Schemes:

G1. $\displaystyle\bigvee_{\sigma_i \in \Sigma_i(\Gamma)} play_i(\sigma_i)$

G2. $\neg(play_i(\sigma_i) \wedge play_i(\sigma_i'))$, for $\sigma_i \neq \sigma_i'$

G3. $play_i(\sigma_i) \Leftrightarrow B_i\, play_i(\sigma_i)$

G4. $RAT_i \Leftrightarrow B_i(RAT_i)$

G5. $(play_i(\sigma_i) \wedge RAT_i) \Rightarrow \bigvee_{S \in \mathcal{X}_{\sigma_i}} \bigwedge_{\sigma_{-i} \in \Sigma_{-i}(\Gamma)} \chi_S(\sigma_{-i})$,
where $\mathcal{X}_{\sigma_i}$ is the set of all $S \subseteq \Sigma_{-i}(\Gamma)$ such that there exists a probability measure μ on $\Sigma_{-i}(\Gamma)$ such that σ_i is a best response to μ and $Supp(\mu) = S$

G6. $(play_i(\sigma_i) \wedge \neg RAT_i) \Rightarrow \bigvee_{S \in \mathcal{Y}_{\sigma_i}} \bigwedge_{\sigma_{-i} \in \Sigma_{-i}(\Gamma)} \chi_S(\sigma_{-i})$,
where $\mathcal{Y}_{\sigma_i}$ is the set of all $S \subseteq \Sigma_{-i}(\Gamma)$ such that there exists a probability measure μ on $\Sigma_{-i}(\Gamma)$ such that σ_i is *not* a best response to μ and $Supp(\mu) = S$.

G1–**G4** are straightforward. **G1** and **G2** say that, in each state, a player plays exactly one strategy; **G3** and **G4** say that a player is certain of his strategy and of whether or not he 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 profiles on which he could put a probability that would justify his playing σ_i. **G6** is interpreted analogously. Notice that these axioms do not specify player i's actual belief. Nevertheless, they are all we need to get completeness.

Let $AX^{CB^*}(\Gamma)$ be the axiom system that results by adding **G1**–**G6** to AX^{CB^*}. For expository purposes, in this paper we restrict our attention to finite Γ-structures.

THEOREM 4. $AX^{CB^*}(\Gamma)$ *is a sound and complete axiomatization of the language* $\mathcal{L}_n^{CB^*}(\Gamma)$ *with respect to the class of all finite* Γ*-structures.*

It is worth noting that this result can be extended to the infinite case provided we take a little more care in defining Γ-structures; more specifically, in the infinite case it becomes important to insist that the strategy function respects the topological structure of Ω.

5. CHARACTERIZING STRONG RATIONALIZABILITY

A strategy σ_i for player i in game Γ is *strongly rationalizable* if, for each player j, there is a set $Z_j \subseteq \Sigma_j(\Gamma)$ and, for each strategy $\sigma_j' \in Z_j$, a probability measure $\mu_{\sigma_j'}$ on $\Sigma_{-j}(\Gamma)$ such that

(a) $\sigma_i \in Z_i$,

(b) $Supp(\mu_{\sigma_j'}) \subseteq Z_{-j}$,

(c) σ_j' is a best response to (the beliefs) $\mu_{\sigma_j'}$, and

(d) for all players j, h and all strategy profiles $\overrightarrow{\sigma}' \in Z_1 \times \cdots \times Z_n$, if $\mu_{\sigma_j'}(\sigma_{-j}') > 0$, then $\mu_{\sigma_h'}(\sigma_{-h}') > 0$.

The standard definition of a rationalizable strategy can be recovered by omitting the final condition. HP prove the following theorem:

THEOREM 5. *A strategy* σ_i *for player* i *in a game* Γ *is strongly rationalizable if and only if there exists a finite probability structure* M *appropriate for* Γ *and a state* ω *such that* $\mathbf{s}_i(\omega) = \sigma_i$ *and* $(M, \omega) \models CB^*(RAT)$.

Of course, this theorem is proved in their paper using their version of justified belief. We seek a version of this result that holds in our logical setting. Two obstacles arise, each stemming from the fact that the implication

$$\omega \models B_i^* \varphi \Rightarrow \omega \in \mathcal{PR}_i[\omega] \tag{4}$$

is licensed according to the HP definition of B_i^*, provided φ is not valid, but it is not licensed according to our definition of B_i^*, even for refutable φ. As (4) plays a central role in many of the results proved by HP, including the proof of Theorem 5, we are motivated to find a suitable substitute. This leads naturally to an investigation of the conditions under which $\omega \models B_i^* \varphi$ even when $\omega \notin \mathcal{PR}_i[\omega]$.

Proposition 1(c) tells us that if $\varphi = B_i \psi$ then $B_i^* \varphi$ might hold even when $\omega \notin \mathcal{PR}_i[\omega]$. However, this particular way in

which (4) can fail is more a feature of our system than a bug, and it will not impede our ability to prove Theorem 5 using our definition of B_i^*. A more serious problem is that $B_i^* \varphi$ holds trivially not just when φ is valid, but, as Proposition 1(b) shows, whenever φ is true at all states in the structure. Thus, in order for $B_i^* \varphi$ to have any bite at all, it is not enough that φ be merely refutable, but φ must in fact be refuted at some state in the structure. This is the first of the two obstacles mentioned above: if $B_i^* \varphi$ is to imply anything at all, the structure must be "sufficiently rich" as to refute φ, if φ is refutable at all.

The second obstacle is perhaps best observed by recalling that if M is a structure in which R_i is the identity, then $M \models B_i^* \varphi \Leftrightarrow (B_i \varphi \wedge \varphi)$. Since we could very well have $\omega \models B_i \varphi$ and $\omega \models \varphi$ without having $\omega \in \mathcal{PR}_i[\omega]$, this suggests that a second type of "richness" property is required, one which ensures that the relations R_i are "sufficiently large".

We capture the notion of "sufficiently large" using the notion of a maximal accessibility relation, as defined in Section 3. We call a Γ-structure *maximal* if its accessibility relations are maximal in this sense.

PROPOSITION 4. *Let* $(\Omega, \mathbf{s}, \overrightarrow{\mathcal{PR}}, \overrightarrow{R})$ *be a maximal* Γ*-structure. For all* $\omega \in \Omega$, *if* $\omega \models B_i^* \varphi$ *and there exists* $\omega' \in \Omega$ *such that* $\mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega)$ *and* $\omega' \models \neg \varphi$, *then* $\omega \in \mathcal{PR}_i[\omega]$.

PROOF. If not, then conditions (a) and (b) of the definition given in (3) are satisfied and thus $\omega R_i \omega'$; this contradicts the assumption that $\omega \models B_i^* \varphi$, since $\omega' \models \neg \varphi$. $\square$

Thus, maximal Γ-structures license the implication in (4), provided that the Γ-structure refutes φ at some state ω' with $\mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega)$. The following definition and lemma will therefore provide exactly the tools we need to formulate and prove our version of Theorem 5.

Fix a Γ-structure $M = (\Omega, \mathbf{s}, \overrightarrow{\mathcal{PR}}, \overrightarrow{R})$. A state $\omega \in \Omega$ is called *i-rich with respect to φ (in M)* if there exists $\omega' \in \Omega$ such that $\mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega)$ and $(M, \omega') \not\models \varphi$. More generally, given $S \subseteq \Omega$ and $\mathcal{F} \subseteq \mathcal{L}_n^{CB^*}(\Gamma)$, say that *$S$ is i-rich with respect to $\mathcal{F}$* if for all $\omega \in S$ and all $\varphi \in \mathcal{F}$, ω is i-rich with respect to φ.

LEMMA 1. *Every finite* Γ*-structure* $(\Omega, \mathbf{s}, \overrightarrow{\mathcal{PR}})$ *can be extended to a finite* Γ*-structure* $(\Omega', \mathbf{s}', \overrightarrow{\mathcal{PR}'})$ *such that, for each* $1 \leq i \leq n$, Ω *is i-rich with respect to* $\{RAT_j : j \neq i\}$ *in* $(\Omega', \mathbf{s}', \overrightarrow{\mathcal{PR}'})$.

PROOF. See the full paper. $\square$

THEOREM 6. *A strategy* σ_i *for player* i *in a game* Γ *is strongly rationalizable if and only if there exists a finite maximal* Γ*-structure* M *such that*

(a) *for each* $1 \leq i \leq n$, $[\![CB^*(RAT)]\!]_M$ *is i-rich with respect to* $\{RAT_j : j \neq i\}$, *and*

(b) *there is some* $\omega \in \Omega$ *such that* $\mathbf{s}_i(\omega) = \sigma_i$ *and* $(M, \omega) \models CB^*(RAT)$.

PROOF. Proposition 4 together with Lemma 1 can be used to change the proof of Theorem 5 given by HP into a proof of this theorem. We defer details to the full paper. $\square$

6. MAXIMALITY AND RICHNESS REVISTED

Theorem 6, although analogous to Theorem 5 insofar as it establishes a correspondence between strong rationalizability and CJBR, loses some of its force due to the additional requirements of maximality and richness. If the goal was to build a logic for which *consistency* of the formula $play_i(\sigma_i) \wedge CB^*(RAT)$ is equivalent to strong rationalizability, then Theorem 6 as it stands has fallen short of that goal. One approach to getting such a logic would be to get axioms that force the maximality and i-richess requirements of Theorem 6. Axiomatizing maximality is accomplished (Theorem 2), but axiomatizing i-richness proves elusive.

But perhaps we do not have to go quite this far. Maximality and i-richness are sufficient requirements on a Γ-structure to ensure that it satisfies the implication (4). Although (4) is a key characteristic of the HP definition of justified belief, and is used in the HP proof of Theorem 5, a close examination of the proof reveals that it is not essential. It can be replaced by a syntactic requirement.

Let $\mathcal{A}$ denote the collection of all formulas of the form

$$CB^*(RAT) \Rightarrow (B_i \neg play(\overrightarrow{\sigma}) \Rightarrow \neg play(\overrightarrow{\sigma})),$$

where $1 \leq i \leq n$ and $\overrightarrow{\sigma} \in \Sigma(\Gamma)$. The structure of these formulas is reminiscent of (1); in fact, each element of $\mathcal{A}$ is implied by a formula of the form (1) (since $CB^*(RAT)$ implies $B_i^*(RAT)$). Moreover, Proposition 4 guarantees that any maximal Γ-structure satisfying condition (a) in Theorem 6 validates every formula in $\mathcal{A}$. On the other hand, as already alluded to, modifying the proof of Theorem 5 yields the following result.

PROPOSITION 5. *If M is a finite Γ-structure such that $M \models \mathcal{A}$, and $(M, \omega) \models CB^*(RAT) \wedge play_i(\sigma_i)$, then σ_i is strongly rationalizable.*

From Proposition 5 together with the preceding discussion and Theorem 6 we can then deduce:

THEOREM 7. *A strategy σ_i for player i in a game Γ is strongly rationalizable if and only if there exists a finite Γ-structure M such that $M \models \mathcal{A}$ and a state ω satisfying $\mathbf{s}_i(\omega) = \sigma_i$ and $(M, \omega) \models CB^*(RAT)$.*

This theorem generates several interesting follow-up questions to which we do not yet know the answers. Does $\mathcal{A}$ give an axiomatization of the class of maximal Γ-structures satisfying condition (a) in Theorem 6? We have argued already that it is sound, but completeness remains an open question, despite the suggestive juxtaposition of Theorems 6 and 7. In either case, what is the relationship between the axioms that characterize maximality and the collection $\mathcal{A}$? Can axioms for i-richness alone be teased out of $\mathcal{A}$ in some fashion? If the class of Γ-structures satisfying $\mathcal{A}$ turns out to be strictly smaller than the class considered in Theorem 6, it is natural to wonder whether we can find an appropriate semantic characterization of the former class.

7. ACKNOWLEDGEMENTS

References

[1] S. Artemov. Explicit provability and constructive semantics. *The Bulletin of Symbolic Logic*, 7(1):1–36, 2001.

[2] S. Artemov and E. Nogina. Introducing justification to epistemic logic. *Journal of Logic and Computation*, 15(6):1059–1073, 2005.

[3] A. Brandenburger and E. Dekel. Rationalizability and correlated equilibria. *Econometrica*, 55:1391–1402, 1987.

[4] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge*. MIT Press, Cambridge, Mass., 1995. A slightly revised paperback version was published in 2003.

[5] M. Fitting. A quantified logic of evidence. *Annals of Pure and Applied Logic*, 152(1-3):67–83, 2008.

[6] J. Y. Halpern and G. Lakemeyer. Multi-agent only knowing. *Journal of Logic and Computation*, 11(1):41–70, 2001.

[7] J. Y. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379, 1992.

[8] J. Y. Halpern and R. Pass. Justified belief and rationality. Unpublished manuscript; available at www.cs.cornell.edu/home/halpern/papers/cbr.pdf, 2011.

[9] D. G. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52(4):1029–1050, 1984.

[10] R. W. Rosenthal. Games of perfect information, predatory pricing, and the chain store paradox. *Journal of Economic Theory*, 25:92–100, 1982.

[11] T. Tan and S. Werlang. The Bayesian foundation of solution concepts of games. *Journal of Economic Theory*, 45(45):370–391, 1988.