An Axiomatic Approach to the Logical Omniscience Problem

Barton L. Lipman
Department of Economics
Queen's University
Kingston, Ontario K7L 3N6
email: lipmanb@qed.econ.queensu.ca

Abstract

Standard models of knowledge have the unrealistic property that agents are logically omniscient in the sense that they know all logical implications of their information. While many nonstandard logics have been proposed to avoid this problem, none has an obvious claim as the "right" logic to use. I show how to derive such a logic as part of a representation of an agent's preferences. In this sense, the agent's logic is given the same basis as a utility function or subjective probability. I provide necessary and sufficient conditions for a given logic to be part of a representation of preferences. Unfortunately, the conditions are not easily interprettable in general. To illustrate them further, I summarize some of my earlier results (Lipman [1993a]) on when the agent's logic is a version of the logic of inconsistency proposed by Rescher and Brandom [1979]. I also discuss the difficulties of representing an agent as using Levesque's logic of implicit belief (Levesque [1984]) or some form of resource–bounded computation.

1 Introduction

It has long been known that the standard possible worlds approach to representing knowledge and beliefs has one very important implication, dubbed by Hintikka [1975] the problem of *logical omniscience*. The possible worlds approach says that an agent knows that φ is true if and only if φ is true in every world the agent conceives of as possible. Suppose the agent learns that φ is true where $\varphi \rightarrow \psi$ is a tautology. If every world the agent conceives of as possible is logically consistent, then $\varphi \rightarrow \psi$ must be true in every such world. Hence in any such world, if φ is true, ψ is true as well. Therefore, an agent who learns that φ is true *must* recognize that ψ is true. In this sense, the agent knows every logical implication of his knowledge. While this is a very attractive property for the study of ideal reasoners, it is unpalatable as an assumption about real people.

I believe that game theorists should also be very interested in relaxing the logical omniscience assumption. Many of the examples said to be "paradoxical," such as the centipede game (Rosenthal [1981], Reny [1986], Binmore [1987]) or van Damme's [1989] dollar–burning example,

rely on a complex deduction from a simple and plausible set of hypotheses. The paradox arises because we believe that the hypotheses may well be known to a real agent, but we are reluctant to believe that a real agent would reach the conclusion. It is precisely the assumption of logical omniscience which makes this view difficult to formalize in standard models.

Fortunately, there is a simple — even obvious — solution to the problem. If some of the worlds the agent conceives of as possible are *not* logically consistent, then the chain of reasoning above is broken. If the agent conceives of a world in which $\varphi \rightarrow \psi$ is true, φ is true, but ψ is false, then learning φ does not lead the agent to recognize that ψ is true, even if he already knows that $\varphi \rightarrow \psi$ is true. I will refer to such worlds interchangably as *nonstandard possible worlds* (following Rescher and Brandom [1979]) or *impossible possible worlds* (following Hintikka [1975]).

The difficulty with this solution, unfortunately, is also quite obvious: what should we assume about the impossible possible worlds? Put differently, exactly which nonstandard logic should we use to describe the reasoning of real agents? It is quite clear what "perfect reasoning" entails; it is not at all obvious how to give a precise formulation of "imperfect reasoning."

In this paper, I propose an approach to this problem (see also Lipman [1992, 1993a]). The idea is to derive the agent's "logic" by analyzing his preferences. In a sense, then, the agent's logic is derived as a representation of preferences in the same way a utility function or subjective probabilities would be derived. Intuitively, if the agent's reasoning does indeed affect his choices, this effect must be observable in some fashion. The natural place to look for this effect is the agent's preferences, or, more specifically, the way the agent's preferences vary with his information.

The simplest way to see this clearly is to suppose that φ and ψ are logically equivalent propositions, yet the agent does not respond to these pieces of information in the same way. That is, his preferences if he is told that φ is true (and is told nothing about ψ directly) differ from the preferences he has if he is told that ψ is true (and is told nothing about φ directly). Then we can infer that the agent does not recognize the fact that φ and ψ are logically equivalent. Put differently, there must be at least one impossible possible world for this agent in which one of the two propositions is true and the other is not.

The rest of this paper is organized as follows. In Section 2, I give the basic framework for relating the agent's logic to his preferences via impossible possible worlds. I also state a simple theorem (proved in Lipman [1993a]) which shows that this approach allows us to "rationalize" virtually any preferences. More precisely, given any preferences satisfying a relatively weak condition, we can represent those preferences as arising from some form of nonstandard reasoning. One way to view this result is that nonstandard logics which are rarely or never part of such a representation of preferences have few or no useful behaviorial implications. As I discuss later, this view has its limitations.

Next I discuss how the characteristics of a nonstandard logic relate to the set of preferences which can be represented as arising from that form of reasoning. I show that there are two key considerations. First, if in the nonstandard logic two propositions are equivalent, then the

agent's preferences must respect this fact if he is to be represented as using this logic.

Second, the preferences in response to certain pieces of information must satisfy special conditions. These pieces of information fall into two categories: propositions which lead the agent to infer perfectly and sets of propositions which are true in the same set of impossible possible worlds. Clearly, the agent's preferences in response to propositions in the former category must be more structured since we cannot use imperfect reasoning to help represent the agent's response. Propositions in the latter category must lead to preferences which are "symmetric" in an unfortunately awkward sense.

In Section 3, I illustrate the use of the framework. First, I use the results of Section 2 to as an alternative approach to some results in Lipman [1993a]. Specifically, I give three logics based on the logic of inconsistency proposed by Rescher and Brandom [1979]. The least restrictive of these turns out to have no propositions in either of the two categories above, so that an agent can be represented as using this logic as long as his preferences respect the logic's notion of equivalence. The more restrictive of the logics do have certain propositions in the two problematic categories, but fortunately are propositions with particularly nice features. As a second illustration, I discuss the difficulties of representing an agent as using Levesque's [1984] logic of implicit belief (as reinterpreted by Fagin, Halpern, and Vardi [1990]). These difficulties are related to the critique of this logic noted by Fagin, Halpern, and Vardi. Finally, I discuss the difficulties of representing an agent as using some form of resource–bounded reasoning. Proofs are omitted for brevity.

Related Literature: There are several papers which bear very strong connections to this work. Gilboa and Schmeidler [1992] show that Choquet expected utility — that is, expected utility with respect to a nonadditive probability measure — is equivalent to expected utility on an enlarged state space. Similarly, it is well-known that Shafer's [1976] belief functions — introduced as an alternative to probability for representing uncertainty — can be derived from additive functions on a larger state set. Both enlargements of the state space can be seen as introducing impossible possible worlds. Also, Morris [1992] provides an axiomatic derivation of nonpartitional information structures which has some similarities to my work. In particular, he also uses the way preferences vary with information to derive statements about the agent's reasoning. For more details on the relationship between Morris' work and my own, see Lipman [1993b].

2 Framework for Analysis

Notational Conventions. For any sets A and B, A^B denotes the set of all functions $f: B \to A$, 2^B the set of all subsets of B, and $\#B$ the cardinality of B. If B is a collection of sets, then $\cap B$ is the intersection of all the sets in B and $\cup B$ is the union of all the sets in B.

To model the way preferences vary with information, I need a model of information which does not presume logical omniscience. Hence I begin with propositions as abstract variables, rather than sets of possible worlds. Let Φ_0 denote the set of atomic propositions. For simplicity,

I assume that Φ_0 is finite and contains at least two elements. The set of all propositions, Φ, is the smallest set containing Φ_0 which is closed under $\neg$, $\vee$, $\wedge$, and $\rightarrow$ (which are "not," "or," "and," and "implies" respectively). That is, if $\varphi \in \Phi$, then $\neg\varphi \in \Phi$ and if $\varphi, \psi \in \Phi$, then $\varphi \vee \psi \in \Phi$, $\varphi \wedge \psi \in \Phi$, and $\varphi \rightarrow \psi \in \Phi$. Note that while Φ_0 is finite, Φ must be infinite. Since I study agents who reason imperfectly, I introduce all of these operations separately rather than relating them according to the usual definitions.

The next ingredient I require is a notion of what "correct" logical deduction is. This is modeled as follows. A state of the world is a collection of propositions which constitutes a complete and logically consistent description of how the world might be. That is, a state is a maximal logically consistent subset of Φ. More formally, $s \subseteq \Phi$ is a *state of the world* or a *possible world* if for all $\varphi, \psi \in \Phi$,

$$\begin{aligned}
\varphi \in s \quad &\text{if and only if} \quad \neg\varphi \notin s \\
\varphi \vee \psi \in s \quad &\text{if and only if} \quad \varphi \in s \text{ or } \psi \in s \\
\varphi \wedge \psi \in s \quad &\text{if and only if} \quad \varphi \in s \text{ and } \psi \in s \\
\varphi \rightarrow \psi \in s \quad &\text{if and only if} \quad \neg\varphi \in s \text{ or } \psi \in s
\end{aligned}$$

Let S denote the set of all possible worlds. A possible world is completely determined by the atomic propositions it contains; hence the finiteness of Φ_0 implies that S is finite. Of course, each element of S contains infinitely many propositions.

Fix a finite number $K \geq 1$. The agent's information will take the form of a set of K or fewer propositions.[1] For any $P \subseteq \Phi$ such that $\#P \leq K$, let $S(P) = \{s \in S \mid P \subseteq s\}$. That is, $S(P)$ is the collection of states of the world in which all the K or fewer propositions in P are true. If $S(P) \neq \emptyset$, then P is called an *information set.* Let $\mathcal{P}$ denote the collection of information sets. For notational ease, I often write the singleton $\{\varphi\}$ as φ.

If $S(\varphi) = S$, then φ is a *tautology.* If $S(\varphi) = \emptyset$, then φ is a *contradiction.* For convenience, I assume that there is a special proposition **true** $\in \Phi$ such that **true** is a tautology and a special proposition **false** $\in \Phi$ such that **false** is a contradiction.

Let X be the set of *consequences.* As in Savage [1954], the interpretation of a consequence is that it is as complete a specification of the outcome of a choice as is necessary to describe an agent's evaluation of that outcome. For simplicity, I will take $X = \mathbf{R}$. Let $F = X^S$ denote the set of *acts.* That is, an act is a function from states into consequences, so a choice is viewed in terms of the relationship it creates between external events (which propositions hold) and consequences.

For each $P \in \mathcal{P}$, we have a preference relation on F, $\succ_P$, to be interpreted as the agent's preferences given information set P. That is, $f \succ_P g$ is interpreted as saying that the agent strictly prefers act f to act g if he learns that all the propositions in the set P were true. Let PREF denote the collection of these preference orderings and let $\succ = \succ_{\textbf{true}}$. I emphasize that I make no assumption about the agent's self-awareness. I assume that we, as modellers, know

[1] Because the agents here may not carry out conjunction properly, a set of propositions is not equivalent to the conjunction of the propositions in the set.

how the agent would respond to each possible piece of information, not that the agent himself knows this *ex ante*. Note also that I do not ask for information about the agent's preferences in response to nonsensical pieces of information such as $\varphi \wedge \neg\varphi$. Finally, I emphasize that these preferences are interpretted as the preferences the agent has *after* whatever deduction he carries out. Hence they should reflect whatever processing of this information he carries out.

A natural way to represent these preferences is with expected utility. Say that an information set $P \in \mathcal{P}$ is *null* if $f \sim_P g$ for all $f, g \in F$.

Definition 1 PREF *is* expected utility representable (EUR) *if there is a function* $u: X \to \mathbf{R}$ *and a probability measure* μ *on* S *such that for all nonnull information sets* $P \in \mathcal{P}$, $\mu(S(P)) > 0$ *and*

$$f \succ_P g \quad \textit{if and only if} \quad \mathrm{E}_\mu[u(f(s)) \mid s \in S(P)] > \mathrm{E}_\mu[u(g(s)) \mid s \in S(P)],$$

where $\mathrm{E}_\mu[\cdot \mid s \in S(P)]$ *denotes the expectation with respect to the measure* μ *conditional on the event* $s \in S(P)$.

It is straightforward to restate the Savage [1954] axioms in this framework to give sufficient conditions for such a representation.[2]

There is one necessary condition for an expected utility representation which is implicit in the usual framework and so is not normally discussed. Say that P and Q are *logically equivalent* if $S(P) = S(Q)$.

Definition 2 PREF *is* equivalence respecting (ER) *if for all information sets* P *and* Q, $S(P) = S(Q)$ *implies* $\succ_P = \succ_Q$.

Clearly, if PREF is EUR, then for any logically equivalent P and Q, expected utility conditional on $s \in S(P)$ must be the same as expected utility conditional on $s \in S(Q)$. Hence we must have $\succ_P = \succ_Q$, so PREF satisfies ER.

ER is an implausible assumption because agents are unlikely to always recognize logical equivalence. The problem is not that the agent knows that $S(P) = S(Q)$ and wishes to behave differently when receiving information P than when receiving Q. Instead, the agent simply doesn't realize that $S(P) = S(Q)$. This suggests the following approach.

Let $\mathcal{S}$ denote the set of all subsets of 2^Φ. Given any $S' \in \mathcal{S}$ and $P \subseteq \Phi$, let $S'(P) = \{s \in S' \mid P \subseteq s\}$. Let $\mathcal{S}^*$ denote the set of $S' \in \mathcal{S}$ such that $S \subseteq S'$, $S'(\mathsf{true}) = S'$, and $S'(\mathsf{false}) = \emptyset$.

Definition 3 PREF *is* extended expected utility representable (XEUR) *if there exist*

1. *a set* $S^* \in \mathcal{S}^*$;

[2]The finiteness of the state space does complicate matters. See Gul [1992] and Chew and Karni [1992].

2. *a function* h: $F \to X^{S^*}$ *with* $h(f)(s) = f(s)$ *for all* $s \in S$;

3. *a function* u: $X \to \mathbf{R}$ *and a probability measure* μ *on* S^* *such that* $\mu(S(P)) > 0$ *for all nonnull information sets* P

where for all $f,\ g \in F$,

$$f \succ_P g \quad \textit{if and only if} \quad \mathrm{E}_\mu[u(h(f)(s)) \mid s \in S^*(P)] > \mathrm{E}_\mu[u(h(g)(s)) \mid s \in S^*(P)].$$

Call (S^*, h, u, μ) *an* XEU representation *of* PREF *and* S^* *a* part of *an XEU representation.*

In other words, PREF is XEUR if we can extend the state set from S to S^* — *i.e.*, introduce impossible possible worlds — and extend all acts to the new state set (via the function h) in such a way that the preferences are represented by expected utility on the larger state set.

The following is an obviously necessary condition for PREF to be XEUR.

Definition 4 PREF *is* representable (REP) *if for every nonnull information sets* P, *there exists a function* u_P: $F \to \mathbf{R}$ *such that*

$$f \succ_P g \quad \textit{if and only if} \quad u_P(f) > u_P(g).$$

Obviously, when PREF is not REP, some nonnull $P \in \mathcal{P}$ has no utility function, and, consequently, it is not XEUR. Necessary and sufficient conditions for REP are well known. Note, however, that REP is vastly weaker than EUR.

The following theorem follows from Lipman [1993a].

Theorem 1 *If* PREF *is REP and has finitely many distinct elements, then it is XEUR.*

Theorem 1 implies that large class of preferences can be represented by some kind of imperfect reasoning. Given this, one could argue that nonstandard logics which cannot be used to represent preferences this way are of questionable usefulness in predicting behavior. Later, though, I will point out a limitation on this interpretation.

Assume a nonstandard logic in the form of a set of possible worlds $S^* \in \mathcal{S}^*$. We wish to know the set of preferences such that an agent can be represented as using this logic. Unfortunately, the conditions I give to answer this refer to preferences only in an indirect fashion. The implications for preferences are easy to see for some logics but not so easy for others.

Definition 5 *Two collections of information sets* $\mathcal{P}_1,\ \mathcal{P}_2 \subseteq \mathcal{P}$ *are* partitionally equivalent with respect to $S' \in \mathcal{S}^*$ *if*

1. $S'(P) \setminus S$ *is nonempty for every* $P \in \mathcal{P}_1 \cup \mathcal{P}_2$

2. *the sets* $\{S'(P) \setminus S\}_{P \in \mathcal{P}_i}$ *are mutually disjoint for* $i = 1, 2$

3. $\bigcup_{P \in \mathcal{P}_1} S'(P) \setminus S = \bigcup_{P \in \mathcal{P}_2} S'(P) \setminus S$.

Note that if $S^*(P) = S^*(Q)$ and $S^*(P) \setminus S \neq \emptyset$, then $\{P\}$ and $\{Q\}$ are partitionally equivalent with respect to S^*.

Theorem 2 *Let* $S^* \in \mathcal{S}^*$ *be part of an XEU representation of* PREF. *Then there exist:*

1. *a function* $u: X \to \mathbf{R}$

2. *a probability measure* μ *on* S^* *such that* $\mu(S(P)) > 0$ *for all nonnull information sets* P

3. *for each nonnull information set* P, *a function* $u_P: F \to \mathbf{R}$ *representing* $\succ_P$

such that

C1 *For all nonnull* P *with* $S^*(P) = S(P)$,

$$f \succ_P g \text{ if and only if } \mathrm{E}_\mu[u(f(s)) \mid s \in S(P)] > \mathrm{E}_\mu[u(g(s)) \mid s \in S(P)].$$

C2 *For all* $\mathcal{P}_1, \mathcal{P}_2 \subseteq \mathcal{P}$ *such that* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *are partitionally equivalent with respect to* S^*, *for all* $f \in F$,

$$\sum_{P \in \mathcal{P}_1} \left[u_P(f) - \sum_{s \in S(P)} \mu(s) u(f(s)) \right] = \sum_{P \in \mathcal{P}_2} \left[u_P(f) - \sum_{s \in S(P)} \mu(s) u(f(s)) \right].$$

Moreover, if S^* *is finite and* u *is onto* $\mathbf{R}$ *then* S^* *is part of an XEU representation of* PREF.

A few remarks on this theorem are needed. First, the necessary condition of Theorem 2 requires existence of utility functions representing the individual $\succ_P$ orders and so, unsurprisingly, requires REP.

Second, the necessity of condition (C1) is clear: given any information set which leads the agent to rule out all impossible worlds, imperfect reasoning cannot help explain preferences.

Third, condition (C2) is more intuitive than it may appear. For example, suppose $S^*(P) = S^*(Q)$, so that $\{P\}$ and $\{Q\}$ are partitionally equivalent with respect to S^*. Then (C2) implies that for all $f \in F$,

$$u_P(f) - \sum_{s \in S(P)} \mu(s) u(f(s)) = u_Q(f) - \sum_{s \in S(Q)} \mu(s) u(f(s)).$$

But if $S^*(P) = S^*(Q)$, then $S(P) = S(Q)$, so this implies $u_P(f) = u_Q(f)$ for all f — that is, $\succ_P = \succ_Q$. Hence (C2) implies that the agent recognizes those logical equivalences preserved by the nonstandard logic.

The necessity of (C2) is easily proven. Suppose we have an XEU representation (S^*, h, u, μ). For each information set P, define

$$u_P(f) = \sum_{s \in S^*(P)} \mu(s) u(h(f)(s)).$$

Clearly, u_P must represent $\succ_P$. Note that

$$u_P(f) = \sum_{s \in S(P)} \mu(s) u(f(s)) + \sum_{s \in S^*(P) \setminus S} \mu(s) u(h(f)(s))$$

so

$$(1) \qquad \sum_{s \in S^*(P) \setminus S} \mu(s) u(h(f)(s)) = u_P(f) - \sum_{s \in S(P)} \mu(s) u(f(s)).$$

Consider any partitionally equivalent $\mathcal{P}_1$ and $\mathcal{P}_2$ and let $\hat{S} = \cup_{P \in \mathcal{P}_1} S^*(P) \setminus S$. Clearly,

$$\sum_{s \in \hat{S}} \mu(s) u(h(f(s)) = \sum_{P \in \mathcal{P}_1} \left[\sum_{s \in S^*(P) \setminus S} \mu(s) u(h(f)(s)) \right] = \sum_{P \in \mathcal{P}_2} \left[\sum_{s \in S^*(P) \setminus S} \mu(s) u(h(f)(s)) \right].$$

Substituting from (2) yields condition (C2).

While the conditions of this theorem do not immediately convey a great deal of intuition, the result makes it clear how to check what preferences can be represented by certain logics. For example, it provides a very quick way to derive the results shown in Lipman [1993a] regarding Rescher and Brandom's [1979] logic of inconsistency.

3 The Agent's Logic

Theorems 1 and 2 focus on the set of worlds the agent considers possible. It is often more intuitive to interpret such a set in terms of the inference rules it generates. Given any $S' \in \mathcal{S}^*$, define a relation on sets of propositions P, $Q \subseteq \Phi$ by

$$P \overset{S'}{\hookrightarrow} Q \text{ if and only if } \bigcap_{\varphi \in P} S'(\varphi) \subseteq \bigcup_{\psi \in Q} S'(\psi).$$

That is, if S' is the set of worlds the agent considers possible and $P \overset{S'}{\hookrightarrow} Q$, then an agent learning that all the propositions in P are true will infer that at least one of the propositions in Q is true. Let $\hookrightarrow$ denote $\overset{S}{\hookrightarrow}$.

Below are some properties one may the agent's inference rule to satisfy.

Definition 6 $S^* \in \mathcal{S}^*$ *satisfies* simple inference *(SI) if for all* φ, $\psi \in \Phi$, $\varphi \hookrightarrow \psi$ *implies that* $\varphi \overset{S^*}{\hookrightarrow} \psi$.

Intuitively, if S^* satisfies SI, then an agent learning the single premise φ infers any one conclusion tautologically implied by it. This is similar to the rule of generalization typically used in modal logic. It is not hard to show that SI implies $S^*(\varphi) = S^*$ for all tautologies φ and $S^*(\psi) = \emptyset$ for all contradictions ψ. In this sense, SI requires that the agent knows all tautologies and rules out every contradiction.

Definition 7 *S^* satisfies the* rule of conjunction (ROC) *if for all $\varphi, \psi \in \Phi$,*

1. *φ, $\psi \in P$ implies $P \overset{S^*}{\hookrightarrow} \varphi \wedge \psi$*
2. *$\varphi \wedge \psi \in P$ implies $P \overset{S^*}{\hookrightarrow} \varphi$ and $P \overset{S^*}{\hookrightarrow} \psi$.*

Note that if S^* satisfies SI, then the second condition is redundant.

Definition 8 *S^* satisfies the* rule of disjunction (ROD) *if for all φ, $\psi \in \Phi$,*

1. *$\varphi \in P$ implies $P \overset{S^*}{\hookrightarrow} \varphi \vee \psi$ and $P \overset{S^*}{\hookrightarrow} \psi \vee \varphi$*
2. *$\varphi \vee \psi \in P$ implies $P \overset{S^*}{\hookrightarrow} \{\varphi, \psi\}$.*

Note that if S^* satisfies SI, then the first condition is redundant.

Rescher and Brandom [1979] propose constructing the new worlds in $S^* \setminus S$ from the old ones in S in a particularly simple way: forming unions or intersections of the original states. The simplest way to allow such possibilities is the following. Recall that for any collection of sets B, $\cap B$ denotes the intersection of the sets in B and $\cup B$ denotes the union. Given any $S' \in \mathcal{S}$, let

$$I(S') = \{s^* \subseteq \Phi \mid s^* = \cap B \text{ for some } B \subseteq S'\}$$

and

$$U(S') = \{s^* \subseteq \Phi \mid s^* = \cup B \text{ for some } B \subseteq S'\}.$$

Finally, let $\tau(S')$ denote the smallest topology on $\cup S'$ containing S'. (See Kelly [1955], pp. 46–48.) The natural alternatives to consider are $I(S)$, $U(S)$, and $\tau(S)$, which I will simply denote I, U, and τ respectively. It is easy to show that the finiteness of S implies $\tau = U(I)$.

The following theorem is proved in Lipman [1993a].

Theorem 3

1. *S^* satisfies SI if and only if $S^* \subseteq \tau$.*
2. *S^* satisfies SI and ROC iff $S^* \subseteq I$.*
3. *S^* satisfies SI and ROD iff $S^* \subseteq U$.*

For example, suppose $\Phi_0 = \{p, q\}$ and $S = \{s_1, s_2, s_3, s_4\}$ where

$$p, q \in s_1; \qquad p, \neg q \in s_2, \qquad \neg p, q \in s_3, \qquad \text{and} \quad \neg p, \neg q \in s_4.$$

Example 1 *Suppose that* $S^* = \{s_1, s_2, s_3, s_4, s_1 \cup s_2\}$. *By Theorem 3, SI and ROD hold. However, ROC does not hold since* $S^*(q) \cap S^*(\neg q) = \{s_1 \cup s_2\} \not\subseteq \emptyset = S^*(q \wedge \neg q)$, *so we do* not *have* $\{q, \neg q\} \overset{S^*}{\hookrightarrow} q \wedge \neg q$.

Example 2 *Suppose* $S^* = \{s_1, s_2, s_3, s_4, s_1 \cap s_2\}$. $S^* \subseteq I$ *implies that SI and ROC hold. However, ROD does not hold since* $S^*(q \vee \neg q) = S^* \not\subseteq S = S^*(q) \cup S^*(\neg q)$, *so* $q \vee \neg q \overset{S^*}{\hookrightarrow} \{q, \neg q\}$ *is* not *satisfied.*

Example 3 *Suppose* $S^* = \{s_1, s_2, s_3, s_4, s_2 \cup s_4\}$. *As in Example 1, SI and ROD hold but not ROC. Recall that* $p \rightarrow q \in s$ *whenever* $p \notin s$ *or* $q \in s$. *Hence* $S^*(p \rightarrow q) \cap S^*(p) = \{s_1, s_2 \cup s_4\} \not\subseteq \{s_1, s_3\} = S^*(q)$, *so* $\{p, p \rightarrow q\} \overset{S^*}{\hookrightarrow} q$ *fails to hold. On the other hand,* $S^*(p \wedge (p \rightarrow q)) = \{s_1\}$ *so an agent who* simultaneously *learns* p *and* $p \rightarrow q$ *does infer* q.

Theorem 3 allows characterization of the set of preferences which can be represented by attributing inference rules satisfying these simple criteria to the agent. These results, contained in Theorems 4, 5, and 6, are derived by a series of lemmas.

Lemma 1 *If* $U \subseteq S^*$, *then for all information sets* $P \in \mathcal{P}$, $S^*(P) \neq S(P)$.

Lemma 1 implies that condition (C1) is irrelevant whenever $U \subseteq S^*$.

Say that information sets P and Q are *strongly equivalent*, written $P \leftrightarrow Q$ if for all $\varphi \in P$, there is a $\psi \in Q$ such that $\varphi \hookrightarrow \psi$ and vice versa for all $\psi \in Q$. Clearly, strong equivalence implies logical equivalence but the converse is not true. Also, if P and Q are strongly equivalent, then for any S^* satisfying SI, $P \overset{S^*}{\hookrightarrow} Q$ and conversely.

Lemma 2 *If* U, $I \subseteq S^* \subseteq \tau$, *then* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *are partitionally equivalent only if* $\mathcal{P}_1 = \{P\}$, $\mathcal{P}_2 = \{Q\}$, $P \leftrightarrow Q$, *and* $S^*(P) = S^*(Q)$.

Hence if we let $S^* = \tau$, the fact that U, $I \subseteq \tau$ implies that (C2) is relevant only when $P \leftrightarrow Q$. In this case, $S^*(P) = S^*(Q)$ and $S(P) = S(Q)$. Therefore, (C2) reduces to the following simpler condition.

Definition 9 PREF *is* strong equivalence respecting (SER) *if for all information sets* P *and* Q, $P \leftrightarrow Q$ *implies* $\succ_P = \succ_Q$.

Clearly, SER is weaker than ER since ER requires $\succ_P = \succ_Q$ for more pairs of information sets.

Theorem 4 *There is an $S^* \in \mathcal{S}^*$ satisfying SI which is part of an XEU representation of* PREF *iff* PREF *satisfies REP and SER.*

Even though SER would seem to allow the possibility that the agent correctly recognizes certain equivalences but occasionally makes errors regarding certain simple implications, Theorem 4 implies that this distinction has no behaviorial content. On the other hand, it is possible, at least in principle, that the distinction could be important if one wishes to find a representation that satisfies certain other properties as well.

Lemma 3 *Assume $S^* = I$. Then each of the following holds.*

1. *For every information set P, $S^*(P) = S(P)$ if and only if $\#S(P) \leq 1$.*
2. *Condition (C2) is satisfied if and only if it is satisfied whenever $\mathcal{P}_1 = \{P\}$, $\mathcal{P}_2 = \{Q\}$, and $S^*(P) = S^*(Q)$.*
3. *For every pair of information sets P and Q, $S^*(P) = S^*(Q)$ if and only if $S(P) = S(Q)$.*

Lemma 3 implies that when $S^* = I$, condition (C1) is relevant only when $S(P)$ consists of a single state. In this case, it is not hard to show that (C1) reduces to the following condition.

Definition 10 PREF *satisfies* weak state independence (WSI) *if there exists $u: X \to \mathbf{R}$ such that for all $s \in S$ and every nonnull P such that $S(P) = \{s\}$, $f \succ_P g$ iff $u(f(s)) > u(g(s))$.*

This condition is a substantial weakening of Savage's [1954] state independence axiom, P3, and is much weaker than the usual expected utility conditions.

Lemma 3 also implies that when $S^* = I$, condition (C2) is again only relevant when each partition has only a single element and these two are equivalent in the S^* logic. However, unlike the result with $S^* = \tau$, now two information sets are equivalent in the S^* logic if and only if they are logically equivalent. This yields the following theorem.

Theorem 5 *If there is an $S^* \in \mathcal{S}^*$ satisfying SI and ROC which is part of an XEU representation of* PREF*, then* PREF *is REP, ER, and satisfies WSI. If the u in WSI is onto $\mathbf{R}$, then such an S^* exists.*

Lemma 4 *If $S^* = U$, then*

1. *$\mathcal{P}_1$ and $\mathcal{P}_2$ are partitionally equivalent if and only if $\mathcal{P}_1 = \{P\}$ and $\mathcal{P}_2 = \{Q\}$ where either (a) $S^*(P) = S^*(Q)$ or (b) $\#S \setminus S(P) \leq 1$ and $\#S \setminus S(Q) \leq 1$.*

2. $S^*(P) = S^*(Q)$ *if and only if* $P \leftrightarrow Q$.

By Lemma 1, we know that (C1) is irrelevant when $S^* = U$. Lemma 4 says that (C2) is relevant only in very restricted cases. First, it is relevant when it reduces to SER. The second case, where $\#S \setminus S(P_1) \leq 1$ and $\#S \setminus S(Q_1) \leq 1$ but $S^*(P_1) \neq S^*(Q_1)$, reduces to the following condition.

Definition 11 PREF *satisfies* dual weak state independence (DWSI) *if there exist*

1. u_{true} *representing* $\succ$
2. *a function* $u: X \rightarrow \mathbf{R}$ *and a probability measure* μ *on* S
3. *for each* $s \in S$ *and each information set* P *with* $S(P) = S \setminus \{s\}$, *a function* u_P *representing* $\succ_P$

such that for every $f \in F$, *every* $s \in S$, *and every* $P \in \mathcal{P}$ *with* $S(P) = S \setminus \{s\}$,

$$\mu(s)u(x) \;=\; u_{\mathsf{true}}(f) - u_p(f).$$

Weak state independence requires a function $u(x)$ that represents preferences conditional on a single state. DWSI is the dual to WSI in the sense that it requires a function $\mu(s)u(x)$ representing the *difference* in utility associated with a single state.

Theorem 6 *If there is an* $S^* \in \mathcal{S}^*$ *satisfying SI and ROD which is part of an XEU representation of* PREF, *then* PREF *is REP, SER, and satisfies DWSI. If the* u *in DWSI is onto* $\mathbf{R}$, *then such an* S^* *exists.*

Summarizing, Theorem 4 implies that whenever PREF is REP and SER, we can represent these preferences using some form of Rescher and Brandom's logic. How much more structure we can put on the logic depends on the additional structure of the preferences. REP is the weakest condition I can consider. SER, however, is not at all trivial — it requires the agent to recognize logical equivalence in many complex circumstances. This suggests that a logic which does not require SI might be more useful.

One candidate is Levesque's [1984] logic of implicit belief. This logic seems like a natural alternative to Rescher and Brandom since their logic always assumes SI and may add one or both of ROC and ROD, while Levesque's logic always satisfies ROC and ROD, only adding SI to generate standard logic.

Levesque constructs a set of worlds as follows. (This construction differs from Levesque's, but was shown to be equivalent by Fagin, Halpern, and Vardi [1990].[3]) Let L denote the

[3] My treatment of $\rightarrow$ differs from Levesque's and is equivalent to what Fagin, Halpern, and Vardi [1990] call strong implication.

largest subset of $\mathcal{S}^*$ such that there exists a function $\eta: L \to L$ with $\eta(\eta(s)) = s$ for all $s \in L$ and

$$\begin{aligned} \neg\varphi \in s &\quad \text{if and only if} \quad \varphi \notin \eta(s) \\ \varphi \wedge \psi \in s &\quad \text{if and only if} \quad \varphi \in s \text{ and } \psi \in s \\ \varphi \vee \psi \in s &\quad \text{if and only if} \quad \varphi \in s \text{ or } \psi \in s \\ \varphi \to \psi \in s &\quad \text{if and only if} \quad \varphi \notin s \text{ or } \psi \in s \end{aligned}$$

When $\eta(s) = s$, we have $s \in S$. Otherwise, s is an impossible possible world.

The difficulty with setting $S^* = L$ is that there will be very many information sets at which the antecedents of conditions (C1) and (C2) will hold. This problem arises because learning a tautology can be very informative when $S^* = L$. The following theorem is a slight modification of results in Fagin, Halpern, and Vardi [1990].

Theorem 7 *If $S^* = L$, then*

1. *there exists a tautology $\varphi \in \Phi$ such that for all information sets P, we have $S^*(P \cup \{\varphi\}) = S(P \cup \{\varphi\})$.*
2. *for each pair of information sets P and Q with $S(P) \subseteq S(Q)$, there exists a tautology φ' such that*

$$S^*(P) \setminus S = S^*(Q \cup \{\varphi'\}) \setminus S \neq \emptyset.$$

Since the antecedents of (C1) and (C2) will be satisfied for very large classes of information sets, the set of preferences which can be represented with $S^* = L$ is relatively small.

It is worth noting that the tautology φ referred to in the first statement in Theorem 7 simply says that for any atomic proposition p, exactly one of p or $\neg p$ must hold. Similary, the tautology φ' of the second statement of the theorem says the same thing except only for certain of the atomic propositions. As Fagin, Halpern, and Vardi point out, it seems implausible that such statements would convey information to an agent. Interestingly, precisely this odd aspect of Levesque's logic implies that it is part of a representation of preferences only for special preferences.

Another approach to nonstandard logics is based on some notion of resource–bounded computation, where the agent carries out deduction until he runs out of time, energy, or some other resource. While this notion is very intuitive, it is very unlikely to be part of an XEU representation of preferences. Presumably, the inference rules generated by resource–bounded computation will be nontransitive. That is, there will be φ_1, φ_2, and φ_3 such that the agent infers φ_2 from φ_1, infers φ_3 from φ_2, but fails to infer φ_3 from φ_1 because this takes "too long." But for any $S' \in \mathcal{S}$, $\overset{S'}{\hookrightarrow}$ must be transitive.[4] One interpretation is that resource–bounded computation has no behaviorial implications.

[4] This problem is not a necessary feature of resource–bounded reasoning. For example, if an agent is assumed able to carry out any polynomial time calculation, the implied inference relation is transitive.

Such a conclusion is premature, however. While the h function — which extends the acts to S^* — has largely been in the background here, this part of the representation is crucial. Since it models the way the agent views the available acts, it is hardly an innocuous bit of technicality. Hence it is important to find useful and intuitive restrictions on this function. (Some possibilities are discussed in Lipman [1992].) With such restrictions, it may be necessary to extend the notion of an XEU representation and notions like resource–bounded reasoning may be more relevant to such an extension.

Acknowledgements. I thank Debra Holt for helpful discussions and comments on related work and Lenore Zuck for many detailed suggestions, insightful comments, and constructive criticism. Financial support from the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged. Of course, any errors are my own responsibility.

REFERENCES

Binmore, K., "Modeling Rational Players: Part I," *Economics and Philosophy*, 3, 1987, pp. 179–214.

Chew, S. H., and E. Karni, "Choquet Expected Utility with Finite State Space: Commutativity and Act–Independence," University of California–Irvine working paper, 1992.

Fagin, R., J. Halpern, and M. Vardi, "A Nonstandard Approach to the Logical Omniscience Problem," in R. Parikh, ed., *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Third Conference*, San Mateo: Morgan Kaufmann Publishers, 1990.

Gilboa, I., and D. Schmeidler, "Additive Representations of Non–Additive Measures and the Choquet Integral," Northwestern University working paper, 1992.

Gul, F., "Savage's Theorem with a Finite Number of States," Graduate School of Business, Stanford University working paper, 1992.

Hintikka, J., "Impossible Possible Worlds Vindicated," *Journal of Philosophical Logic*, 4, 1975, pp. 475–484.

Kelly, J., *General Topology*, New York: Springer–Verlag, 1955.

Levesque, H., "A Logic of Implicit and Explicit Belief," *Proceedings of National Conference on Artificial Intelligence (AAAI–84)*, 1984, pp. 198–202.

Lipman, B., "Decision Theory with Impossible Possible Worlds," Queen's University working paper, 1992.

Lipman, B., "Logics for Nonomniscient Agents: An Axiomatic Approach," Queen's University working paper, 1993a.

Lipman, B., "Information Processing and Bounded Rationality: A Survey," Queen's University working paper, 1993b.

Morris, S., "Revising Knowledge: A Decision Theoretic Approach," University of Pennsylvania working paper, 1992.

Reny, P., "Rationality, Common Knowledge, and the Theory of Games," Ph.D. dissertation, Princeton University, 1986.

Rescher, N., and R. Brandom, *The Logic of Inconsistency*, Oxford: Basil Blackwell, 1979.

Rosenthal, R., "Games of Perfect Information, Predatory Pricing, and the Chain–Store Paradox," *Journal of Economic Theory*, 25, 1981, pp. 92–100.

Savage, L. J., *The Foundations of Statistics*, New York: Dover, 1954.

Shafer, G., *A Mathematical Theory of Evidence*, Princeton: Princeton University Press, 1976.

van Damme, E., "Stable Equilibria and Forward Induction," *Journal of Economic Theory*, 48, August 1989, pp. 476–496.