# Hypothetical Knowledge and Counterfactual Reasoning*

Joseph Y. Halpern
Dept. Computer Science
Cornell University
Ithaca, NY 14853
halpern@cs.cornell.edu
http://www.cs.cornell.edu/home/halpern

**Abstract:** Samet introduced a notion of *hypothetical knowledge* and showed how it could be used to capture the type of counterfactual reasoning necessary to force the backwards induction solution in a game of perfect information. He argued that while hypothetical knowledge and the *extended information structures* used to model it bear some resemblance to the way philosophers have used *conditional logic* to model counterfactuals, hypothetical knowledge cannot be reduced to conditional logic together with epistemic logic. Here it is shown that in fact hypothetical knowledge can be captured using the standard counterfactual operator "$>$" and the knowledge operator "$K$", provided that some assumptions are made regarding the interaction between the two. It is argued, however, that these assumptions are unreasonable in general, as are the axioms that follow from them. Some implications for game theory are discussed.

## 1 Introduction

It is well understood by now that counterfactual reasoning plays an important role in analyzing rationality in games. In deciding what to do at a given node, a player must analyze what would have happened had he done something else. (See [Aumann 1995; Binmore 1996] for some recent discussion of the role of counterfactual reasoning in games.) Samet [1994, 1996] introduces a notion of *hypothetical knowledge*, and shows how it could be used to capture the type of counterfactual reasoning necessary to force the backwards induction solution in a game of perfect information. In this paper, I examine hypothetical knowledge, show how it

*The work was supported in part by NSF under grant IRI-96-25901 and by the Air Force Office of Scientific Research under grant F49620-96-1-0323. The full version of this paper will appear in the *International Journal of Game Theory*.

can be derived from more standard notions of counterfactuals, and examine some of Samet's assumptions in more detail.

To capture hypothetical knowledge, Samet uses a binary operator $K(H, E)$, which he usually writes as $K^H(E)$. He suggests it should be read as "had $H$ been the case, I would have known $E$". He requires it to satisfy a number of axioms, including $K^H(E) = K^H(K(E))$—if the agent would have known $E$, then he would have known that he would know $E$—and $\neg K^H(E) = K^H(\neg K(E))$—if the agent would not have known $E$, then he would have known that he would not have known $E$. ($K$ is the standard knowledge operator.) Samet argues in [Samet 1994] that while hypothetical knowledge and the *extended information structures* used to model it bear some resemblance to the logic of counterfactuals studied in the philosophical literature [Lewis 1973; Stalnaker 1968; Stalnaker and Thomason 1970], hypothetical knowledge cannot be reduced to counterfactual logic together with epistemic logic.

As I show here, Samet's operator is better read as "had I considered $H$ possible, then I would have known $E$". This reading suggests that we can then represent $K^H(E)$ as $L(H) > K(E)$, where $>$ is the standard counterfactual operator (so that $H > E$ can be read as "if $H$ were the case, then $E$ would be true"). I show that in fact this can be done, provided that we make some assumptions about the interaction between knowledge and counterfactuals, in order to force Samet's axioms to hold. However, as I show by example, these assumptions are not always reasonable, nor are the axioms that follow from them. In particular, it is not always reasonable for an agent to know about his counterfactual knowledge, because it may depend on features of the world that the agent does not know about.

The rest of this paper is organized as follows. In Section 2, I review Samet's framework, while in Section 3, I review conditional logic, which is the framework used by philosophers to model counterfactuals. In Section 4, I show that by combining conditional logic with standard epistemic logic for reasoning about knowledge, we can indeed capture Samet's axioms, by imposing some (not always reasonable) assumptions on the relationship between knowledge and counterfactuals. In doing so, it also becomes clear that Samet's logic is not expressive enough to capture some facets of reasoning about counterfactuals of interest to game theory. In Section 5, there is some discussion of the relevance of these results to game theory.

## 2 Samet's Model

Traditionally, game theorists have considered *information structures* of the form $(\Omega, \Pi_1, \ldots, \Pi_n)$, where $\Omega$ is a set of possible worlds, and for $i = 1, \ldots, n$, $\Pi_i$ is a partition of $\Omega$ into disjoint subsets. Given a world $\omega \in \Omega$, let $\Pi_i(\omega)$ be the element of $\Pi_i$ that contains $\omega$. We can think of $\Pi_i(\omega)$ as the set of worlds that agent $i$ considers possible when in world $\omega$. Intuitively, this set characterizes his state of mind in $\omega$.

In an information structure, we can define the standard unary *knowledge operators* $K_1, \ldots, K_n$, which map subsets of $\Omega$ to subsets of $\Omega$, as follows:

$$K_i(E) = \{\omega : \Pi_i(\omega) \subseteq E\}. \tag{1}$$

It is easy to see that $K_i(E)$ is a union of cells in $\Pi_i$. Intuitively, $K_i(E)$ is the event that agent $i$ knows $E$. For convenience, we define $L_i(E)$ to be an abbreviation for $\neg K_i(\neg E)$, where $\neg$ denotes complementation. Intuitively, $L_i(E)$ is the event "agent $i$ considers $E$ possible". It is well known (and easy to check) that the $K_i$ operator satisfies the following axioms:

A1. $K_i(E) = K_i(K_i(E))$ (*positive introspection*)

A2. $\neg K_i(E) = K_i(\neg K_i(E))$ (*negative introspection*)

A3. For any index set $J$ and events $E_j, j \in J$, $\cap_{j \in J} K_i(E_j) = K_i(\cap_{j \in J} E_j)$ (*intersection*)[1]

A4. $K_i(E) \subseteq E$ (*veridicality*)

Samet considers what he calls *extended information structures*, which are tuples of the form $(\Omega, \Pi_1, \ldots, \Pi_n, T_1, \ldots, T_n)$, where $(\Omega, \Pi_1, \ldots, \Pi_n)$ is an information structure and $T_i : \Pi_i \times (2^\Omega - \emptyset) \rightarrow \Pi_i$. $T_i$ is what Samet calls an *hypothesis transformation.* Given a nonempty event $H \subseteq \Omega$ and a state of mind $P \in \Pi_i$, Samet suggests that we can think of $T_i(P, H)$ as describing what agent $i$'s state of mind would have been had $H$ been the case. As we shall see in Section 4, a more appropriate interpretation is that $T_i(P, H)$ describes what agent $i$'s state of mind would have been had he considered $H$ possible. Samet assumes that hypothesis transformations satisfy the following two properties:

T1. $T_i(P, H) \cap H \neq \emptyset$

T2. If $P \cap H \neq \emptyset$, then $T_i(P, H) = P$.

Samet defines binary *hypothetical knowledge operators* $K_i : (2^\Omega - \emptyset) \times 2^\Omega, i = 1, \ldots, n$, that map events to events. Intuitively, $K_i(H, E)$, usually written $K_i^H(E)$, is meant to represent the event that $i$ would know $E$ had the hypothesis $H$ been true. This is captured by defining

$$K_i^H(E) = \cup\{P \in \Pi_i : T_i(P, H) \subseteq E\}. \tag{2}$$

It is easy to see that the unary $K_i$ operator is equivalent to $K_i^\Omega$, given T1 and T2.

**Lemma 2.1:** *If $T$ satisfies T1 and T2, then $K_i = K_i^\Omega$.*

**Proof:** By T2, for any cell $P \in \Pi_i$, we have that $T_i(P, \Omega) = P$. By Definition (1), $P \subseteq K_i(E)$ iff $P \subseteq E$. On the other hand, since $T_i(P, \Omega) = P$, it follows from Definition (2) that $P \subseteq K_i^H(E)$ iff $P \subseteq E$. The result follows. ▮

Samet wants hypothetical knowledge to satisfy the following axioms, which hold for all events $H$ and $E$:

[1] If $J = \emptyset$, we take the intersection over the empty set to be $\Omega$, as usual. Thus, as a special case of this axiom, we get $\Omega = K_i(\Omega)$.

K1. $K_i^H(E) = K_i(K_i^H(E))$

K2. $K_i^H(E) = K_i^H(K_i(E))$

K3. $\neg K_i^H(E) = K_i^H(\neg K_i(E))$

K4. $\cap_{j \in J} K_i^H(E_j) = K_i^H(\cap_{j \in J} E_j)$, for every index set $J$ and events $E_j, j \in J$

K5. $K_i^H(L_i(H)) = \Omega$

K6. $L_i(H) \cap K_i^H(E) = L_i(H) \cap K_i(E)$

Note that K1–K3 are generalizations of the introspective properties for knowledge, while K4 is a generalization of the intersection property for knowledge.

The following result shows that these axioms uniquely characterize the hypothetical knowledge operator. Moreover, it shows that K5 and K6 correspond to assumptions T1 and T2, respectively.

**Theorem 2.2:** [Samet 1994; Samet 1996] *Let $\mathcal{T}$ be a (possibly empty) subset of $\{T1,T2\}$, let $\mathcal{K}$ be the corresponding subset of $\{K5,K6\}$, and let $(\Omega, \Pi_1, \ldots, \Pi_n)$ be an information structure. All hypothesis transformations $T_1, \ldots, T_n$ on this information structure satisfying the assumptions in $\mathcal{T}$ define hypothetical knowledge operators that satisfy K1–K4 and the axioms in $\mathcal{K}$. Moreover, if $K : (2^\Omega - \emptyset) \times 2^\Omega \rightarrow 2^\Omega$ is a binary operator satisfying K1–K4 and the axioms in $\mathcal{K}$ (for agent i), then there is an hypothesis transformation $T_i$ on $\Pi_i$ satisfying $\mathcal{T}$ such that $K(H,E) = \cup\{P \in \Pi_i | T_i(P,H) \subseteq E\}$.*

As I said in the introduction, one of the goals of this paper is to examine Samet's assumption. The analysis suggests that K3, in particular, is problematic. The proof of Theorem 2.2 shows that K3 follows from the assumption that $T$ is a *function*; that is, it returns a single cell given its two arguments. (See also Theorem 4.5.) The following example, which illustrates the problem with K3 and the assumption that $T$ is a function, was inspired by the later analysis, but can be understood without going through it.

**Example 2.3:** Suppose that agent 1 is in a dark room. He knows that the door is painted either red or blue, but does not know which (and cannot tell since the room is dark). What state of mind will the agent be in if the light is on? He will clearly know either that the door is red or that it is blue, but which he knows depends on the actual situation, which is something the agent does not know. Nor will any amount of introspection will help him figure it out.

Formally, we can model this situation by an information structure consisting of four worlds, $\{(red, off), (blue, off), (red, on), (blue, on)\}$, with the obvious interpretations. (For example, in the world $(red, on)$, the door is red and the light is off.) Let *RED*, *ON*, and *OFF* be the events that the door is red, the light is on, and the light is off, respectively (so that, for example, $OFF = \{(red, off), (blue, off)\}$). The agent's partition consists of three cells: *OFF*, $\{(red, on)\}$, and $\{(blue, on)\}$. Clearly, the agent is only uncertain about the door color if the light is off. Now what should $T_1(\mathit{OFF}, \mathit{ON})$ be? By T1, it must be one of $\{(\mathit{red}, \mathit{on})\}$ or $\{(\mathit{blue}, \mathit{on})\}$, but there is no obvious reason to choose one over the other. The agent does not know what his state of mind would be if the light were on, because he does not know the door color.

This translates to a problem with K3: Clearly $(\mathit{red}, \mathit{off}) \notin K_1^{ON}(\mathit{RED})$: the agent does not know in world $(\mathit{red}, \mathit{off})$ that if the light were on then the door would be red; he considers it possible that it might be blue. On the other hand, we also have $(\mathit{red}, \mathit{off}) \notin K_1^{ON}(\neg K_1(\mathit{RED}))$: the agent does not know if the light were on that he would not know that the door would be red. If the door were actually red, he would in fact know it if the light were on. ∎

# 3 Conditional logic

The more traditional way in the philosophical literature to capture hypothetical or counterfactual reasoning is by means of conditional logic. (See [Stalnaker 1992] for a short and readable survey.) A counterfactual of the form $H > E$, read "if $H$ were the case, then $E$ would be true" is taken to be true in world $\omega$ if at the closest worlds where $H$ is true, $E$ is also true.

To make this precise, we need a "closeness" relation. The original approach, due to Stalnaker and Thomason [Stalnaker 1968; Stalnaker and Thomason 1970], assumes that there is a *selection function* $f : \Omega \times 2^{\Omega} \rightarrow \Omega$; intuitively, $f(\omega, E)$ is the world closest to $\omega$ that satisfies $E$. This implicitly assumes that there is a unique world closest to $\omega$ that satisfies $E$. Many later authors argued that there is not in general a unique closest world; ties should be allowed (see, in particular, [Lewis 1973, pp. 77–81]). I follow this more general interpretation here and take a *counterfactual structure* to be a pair $(\Omega, f)$, where $f : \Omega \times 2^{\Omega} \rightarrow 2^{\Omega}$, although R4 below restricts to the case considered by Stalnaker. Define the binary operator $>$: $2^{\Omega} \times 2^{\Omega} \rightarrow 2^{\Omega}$ in counterfactual structures as follows:

$$H > E = \{\omega : f(\omega, H) \subseteq E\}. \qquad (3)$$

This captures the intuition that $\omega \in H > E$ if the closest worlds to $\omega$ where $H$ is true all satisfy $E$. To simplify the exposition, if $\omega \in H$, we say that $\omega$ is an $H$*-world.*

Given our description of the selection function, the following restrictions on it seem reasonable:

R1. $f(\omega, H) \subseteq H$: the worlds closest to $\omega$ satisfying $H$ are in fact $H$-worlds.

R2. If $H \neq \emptyset$ then $f(\omega, H) \neq \emptyset$: this says that there always is some world closest to $H$ if $H$ is nonempty.

R3. If $\omega \in H$, then $f(\omega, H) = \{\omega\}$: if $\omega$ is an $H$-world, then it is the closest $H$-world to $\omega$.

As I mentioned above, we can also consider a restriction that forces there always to be a *unique* closest world, as was done by Stalnaker.

R4. If $H \neq \emptyset$ then $f(\omega, H)$ is a singleton.

Of course, R4 implies R2. Note that R4 is similar in spirit to Samet's assumption that the hypothesis transformation is a function. As we shall see, there is more than a spiritual similarity between the two assumptions. Once the appropriate connections are made between Samet's approach and standard conditional logic, R4 implies the functionality of hypothesis transformations.

Each of these restrictions corresponds to an axiom. Consider the following axioms:

C0. $\cap_{j \in J}(H > E_j) = H > \cap_{j \in J} E_j$, for any index set $J$ and events $E_j, j \in J$

C1. $(H > H) = \Omega$

C2. $H > \emptyset = \emptyset$ if $H \neq \emptyset$

C3. $H \cap (H > E) = H \cap E$

C4. $H > \neg E = \neg(H > E)$ if $H \neq \emptyset$

**Theorem 3.1:** *Let $S$ be a (possibly empty) subset of $\{R1,R2,R3,R4\}$, let $C$ be the corresponding subset of $\{C1,C2,C3,C4\}$, and let $\Omega$ be a set of worlds. If $f$ is a selection function on $\Omega$ that satisfies the properties in $S$ and $>$ is defined in $(\Omega, f)$ by (3), then $>$ satisfies C0 and the axioms in $C$. Conversely, if $>'$: $\Omega \times 2^{\Omega} \rightarrow 2^{\Omega}$ and satisfies C0 and the axioms in $C$, then there is a selection function $f$ on $\Omega$ satisfying $S$ such that $>'$ is the counterfactual operator $>$ in $(\Omega, f)$.*[2]

**Proof:** It is easy to check that if $f$ satisfies the properties in $S$, then $>$ satisfies C0 and all the properties in $C$. For the second half, given an operator $>'$, define $f(\omega, H) = \cap\{E : \omega \in H >' E\}$. I leave it to the reader to check that $>'$ is the counterfactual operator $>$ in $(\Omega, f)$. ▮

# 4 Conditional epistemic logic

Counterfactuals clearly do not suffice to capture Samet's hypothetical knowledge; we need knowledge as well. Define a *counterfactual information structure* to be a tuple $(\Omega, \Pi_1, \ldots, \Pi_n, f_1, \ldots, f_n)$,

[2] Although completeness results for counterfactuals are well known—the first goes back to Stalnaker and Thomason [1970]—these proofs are syntactic. That is, they start with a language (a collection of formulas) and a notion of what it means for a formula to be true in a counterfactual structure, and then characterize the formulas that are true in all structures. Not surprisingly, there is a close similarity between some of the axioms above and the axioms used to characterize syntactic completeness for conditional logic. For example, C1 and C3 are the derived theorems t4.4 and t4.9 from [Stalnaker and Thomason 1970, p. 31]. Since Stalnaker and Thomason allow $f(\omega, H)$ to be the empty set, they have no analogue of C2. One can derive a finitary analogue of C0 from the standard axioms for counterfactuals given in the literature, but not the infinitary analogue. Indeed, in a precise sense, it is consistent with the standard axioms that the infinitary analogue does not hold; see [Halpern 1998] for discussion of this issue and more details on semantic axiomatizations of counterfactuals.

where $(\Omega, \Pi_1, \ldots, \Pi_n)$ is an information structure and $f_i$, $i = 1, \ldots, n$, is a selection function. In a counterfactual epistemic structure, we can make sense of events defined in terms of both knowledge operators and counterfactual operators. Of course, now the counterfactual operators must be indexed according to the agent, so we have expressions such as $H >_j E$ and $K_i(H >_j E)$. An expression such as $H >_j E$ can be read as "according to agent $j$, if $H$ were the case, then $E$ would be true". Thus, $\omega \in H >_j E$ if, according to the selection function agent $j$ uses at the world $\omega$, the closest worlds to $\omega$ where $H$ is true all satisfy $E$.

Our goal is to find a statement in this framework corresponding to $K_i^H(E)$. To do this, we must first consider carefully how to interpret such a statement. Samet [1996] suggests the reading "Had $H$ been the case, I would have known $E$". The picture he has seems to be the following. Suppose agent $i$ is currently in state of mind $P$ (that is, $P = \Pi_i(\omega)$, where $\omega$ is the state of the world). To evaluate whether the statement "Had $H$ been the case, I would have known $E$", the agent considers what his state of mind would have been if $H$ were true. This is given by $T_i(P, H)$. Since the agent is viewed as having perfect introspection, he can then determine whether $E$ is known in that state of mind. In the state of mind $T_i(P, H)$, $H$ is not necessarily known to be true. The agent realizes that he does not have perfect information, so even had $H$ been true, he might not have known it. However, he would definitely consider it possible. This is the content of T1, which says that $T_i(P, H) \cap H \neq \emptyset$. If $H$ is already considered possible in agent $i$'s current state of mind—that is, if $P \cap H \neq \emptyset$ or, equivalently, if $\omega \in L_i(H)$—then the state of mind that agent $i$ would be in if $H$ were true is his current state of mind. This is the content of T2 (and K6).

In light of this, perhaps a better reading of $K_i^H(E)$ is "if agent $i$ considered $H$ to be possible, then $i$ would have known $E$". We thus capture $K_i^H(E)$ in conditional epistemic logic as $L_i(H) >_i K_i(E)$, abbreviated as $\overline{K}_i^H(E)$. (Another possible translation is considered below.)

It is easy to see that $\overline{K}_i^H$ satisfies K2 and K4, with no assumptions at all on $f_i$. K5 follows easily from C1, since $\overline{K}_i^H(L_i(H))$ becomes $L_i(H) >_i K_i(L_i(H))$, which is equivalent to $L_i(H) >_i L_i(H)$ (by A2 and A4). K6 is a special case of C3. Thus, we get K5 and K6 just by requiring $f_i$ to satisfy the minimal assumptions R1 and R3. If we assume $f_i$ satisfies R2, then we get the additional axiom

K7. $K_i^H(\emptyset) = \emptyset$.

It is easy to see that K7 is a special case of C2. It is not hard to see that it holds in Samet's framework; it follows from K3 and the fact that $K_i(\Omega) = K_i^H(\Omega) = \Omega$. However, in the absence of K3, we have to consider it separately.

To get $\overline{K}_i^H$ to satisfy K1, we need to make some assumptions about the relationship between $f_i$ and $\Pi_i$. Write $\omega \sim_i \omega'$ if $\Pi_i(\omega) = \Pi_i(\omega')$. Intuitively, if $\omega \sim_i \omega'$, then in world $\omega$, the agent considers $\omega'$ possible. We can extend the $\sim_i$ notation to sets by taking $E \sim_i E'$ if, for all $\omega \in E$, there exists some $\omega' \in E'$ such that $\omega \sim_i \omega'$, and for all $\omega' \in E'$, there exists some $\omega \in E$ such that $\omega \sim_i \omega'$.

It might seem reasonable to require that the agent should know his selection function. That is, it might seem reasonable to require

R5$'$. If $\omega \sim_i \omega'$ then $f_i(\omega, H) = f_i(\omega, H)$.

I shall argue in Section 5 that R5$'$ is in fact not always so reasonable. In any case, it turns out that a slightly weaker assumption suffices to get $\overline{K}_i^H$ to satisfy K1, and it is that weaker assumption I focus on now:

R5. If $\omega \sim_i \omega'$ then $f_i(\omega, H) \sim_i f_i(\omega', H)$.

To understand the intuition behind R5, given a selection function $f$, let $\overline{f}(\omega, H) = \cup_{\omega' \in f(\omega,H)} \Pi(\omega')$. Thus, $\overline{f}(\omega, H)$ is the smallest union of cells that contains $f(\omega, H)$. It is easy to see that R5 forces $\overline{f}$ to act the same way on all indistinguishable worlds.

**Lemma 4.1:** *The selection function $f_i$ satisfies R5 if and only if, for all $\omega$, $\omega'$, and $H$, we have that $\omega \sim_i \omega'$ implies $\overline{f}_i(\omega, H) = \overline{f}_i(\omega', H)$.*

Thus, while R5$'$ says that $f_i$ is the same at all worlds the agent considers possible, R5 says that $\overline{f}_i$ is the same at all worlds the agent considers possible. Thus, if $f_i$ satisfies R5, then we can view $\overline{f}_i$ as a function of the cell, not the world. That is, we can write $\overline{f}_i(P, H)$ for $P \in \Pi_i$, taking it to be $\overline{f}_i(\omega, H)$ for some $\omega \in P$. (The choice of $\omega$ does not matter, given R5.) This means that $\overline{f}_i$ is almost an hypothesis transformation. However, $\overline{f}_i(P, H)$ is a union of cells, rather than being a single cell. This is easily seen to suffice for K1; that is, had Samet defined an hypothesis transformation to be a function $T(P, H)$ that returned a union of cells satisfying T1 and T2, then all of his axioms other than K3 would have held. For K3, $T(P, H)$ needs to be a single cell. This is analogous to Stalnaker's requirement that $f(\omega, H)$ return a single world. That is, we can think of $T(P, H)$ as being the unique cell "closest" to $P$ where $L(H)$ holds (i.e., such that $H \cap T(P, H) \neq \emptyset$). It is thus perhaps not surprising that Stalnaker's condition R4, which says that there is a unique world closest to $\omega$ where $H$ holds, gives us K3.[3]

**Theorem 4.2:** *Let $\mathcal{S}$ be a (possibly empty) subset of {R1,R2,R3,R4,R5}, let $\mathcal{K}$ be the corresponding subset of {K5,K7,K6,K3,K1}, and let $(\Omega, \Pi_1, \ldots, \Pi_n)$ be an information structure. If $f_1, \ldots, f_n$ are selection functions such that $(\Omega, \Pi_1, \ldots, \Pi_n, f_1, \ldots, f_n)$ satisfies the properties in $\mathcal{S}$, then all the operators $\overline{K}_i^H$ satisfy K2, K4, and all the axioms in $\mathcal{K}$. Moreover, if $K : (2^\Omega - \emptyset) \times 2^\Omega \longrightarrow 2^\Omega$ is a binary operator satisfying K2, K4, and the axioms in $\mathcal{K}$ (for agent $i$), then there is a selection function $f_i$ on $\Omega$ satisfying the properties in $\mathcal{S}$ such that $K(H, E) = \overline{K}_i^H(E)$.*

[3] Actually, to get K3, it suffices to weaken R4 to require only that $f_i(\omega, H)$ be a subset of some cell in $\Pi_i$. However, this condition is less well-motivated than R4, and all of my later comments still apply if we consider the weaker version.

**Proof:** It is easy to see that if $f_i$ satisfies the properties in $\mathcal{S}$, then $\overline{K}_i^H$ satisfies K2, K4 and all the axioms in $\mathcal{K}$. For the second half, suppose $K$ satisfies K2, K4 and the axioms in $\mathcal{K}$ for agent $i$. There are three cases to consider. If neither K6 nor K3 is in $\mathcal{K}$, then we can define $f_i(\omega, L(H)) = \cap\{E : \omega \in K(H, E)\}$. We leave it to the reader to check that, with this definition, $K(H, E) = \overline{K}_i^H(E)$ and $f_i$ satisfies the properties in $\mathcal{S}$ if its second argument is of the form $L(H)$. We can easily define $f_i(\omega, H')$ if $H'$ is not of the form $L(H)$ for some $H$ (that is, if $H'$ is not the union of cells in $\Pi_i$) so that it satisfies the properties in $\mathcal{S}$. Exactly how it is defined is irrelevant.

This definition must be modified slightly if either K6 or K3 is in $\mathcal{K}$. The trouble is that it follows from K2 that $\cap\{E : \omega \in K(H, E)\}$ is a union of cells in $\Pi_i$. Thus, it will not in general be a singleton and hence will not satisfy R3 or R4. I leave the details of the required modifications to the full paper.

Thus, $\overline{K}_i^H$ corresponds exactly to Samet's $K_i^H$ provided we assume R1–R5. R1–R3 are minimal assumptions for counterfactuals. As we shall see, R4, which says that there is always a unique closest world, is not always reasonable. R5 is even more problematic. However, before discussing these properties, I briefly consider one other way of capturing $K_i^H$ in conditional epistemic logic.

Samet [personal communication, 1997] views $K_i^H$ as an *epistemic* conditional, making statements about an agent's epistemic state. From this point of view, K1 follows almost tautologically. We can capture this intuition by translating $K_i^H(E)$ as $K_i(L_i(H) >_i K_i(E)$, abbreviated as $\overline{\overline{K}}_i^H(E)$.[4] Of course, in the presence of K1, $\overline{K}_i^H(E)$ and $\overline{\overline{K}}_i^H(E)$ are equivalent, although in general they are distinct. It is easy to see that $\overline{\overline{K}}_i^H$ satisfies K1, K2, and K4, with no assumptions at all on $f_i$. Again, to get K5, K6, and K7 correspond to R1, R3, and R2, respectively. Thus, by making the minimal assumptions of counterfactual reasoning, we get all of Samet's properties but K3. We might hope that R4 would give us K3, just as with the previous translation. However, as the following example shows, R4 does not suffice to give us K3.

**Example 4.3** Suppose $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. There is only one agent, and $\Pi_1 = \{\{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4\}\}$. Let $H = \{\omega_3, \omega_4\}$, $f(\omega_1, H) = \{\omega_3\}$, and $f(\omega_2, H) = \{\omega_4\}$. The definition of $f$ for other arguments is irrelevant, so long as it satisfies R1–R4, which can easily be arranged. Note that $f$ does not satisfy R5, since $\omega_1 \sim \omega_2$, but $f(\omega_1) \not\sim f(\omega_2)$. We then have $\omega_1 \in \neg\overline{\overline{K}}^H(\{\omega_3\}) - \overline{\overline{K}}^H(\neg K(\{\omega_3\}))$, so K3 does not hold. ▮

It turns out that to get K3, we need both R4 and R5. Let R45 be the conjunction of R4 and R5.

**Theorem 4.4** *Let $\mathcal{S}$ be a (possibly empty) subset of $\{$R1,R2,R3,R45$\}$, let $\mathcal{K}$ be the corresponding subset of $\{$K5,K7,K6,K3$\}$, and let $(\Omega, \Pi_1, \ldots, \Pi_n)$ be an information structure. If $f_1, \ldots, f_n$ are*

[4] Samet [1994] considers a translation similar to $\overline{\overline{K}}_i^H(E)$—$K_i(H >_i K_i(E))$—and argues that it does not correspond to $K_i^H(E)$.

*selection functions such that* $(\Omega, \Pi_1, \ldots, \Pi_n, f_1, \ldots, f_n)$ *satisfies the properties in* $S$*, then all the operators* $\overline{\overline{K}}_i^H$ *satisfy K1, K2, K4, and all the axioms in* $\mathcal{K}$*. Moreover, if* $K : (2^\Omega - \emptyset) \times 2^\Omega \rightarrow 2^\Omega$ *is a binary operator satisfying K1, K2, K4, and the axioms in* $\mathcal{K}$ *(for agent* $i$*), then there is a selection function* $f_i$ *on* $\Omega$ *satisfying the properties in* $S$ *such that* $K(H, E) = \overline{\overline{K}}_i^H(E)$.

**Proof:** It is easy to check that $\overline{\overline{K}}_i^H$ satisfies K1, K2, K4, and the axioms in $\mathcal{K}$ if $f_i$ satisfies the properties in $\mathcal{S}$. For the second half, we define $f_i$ just as in the proof of Theorem 4.2. I leave it to the reader to check that this has the required properties. ▮

There is an even more direct relationship between $\overline{\overline{K}}_i^H$ and Samet's $K_i^H$ that becomes clear if we consider a slight generalization of Samet's framework. Define a *generalized information structure* to be a tuple of the form $(\Omega, \Pi_1, \ldots, \Pi_n, T_1, \ldots, T_n)$, where $(\Omega, \Pi_1, \ldots, \Pi_n)$ is an information structure and $T_i : \Pi_i \times (2^\Omega - \emptyset) \rightarrow (2^{\Pi_i} - \emptyset)$ is a *generalized hypothesis transformation*. We can think of $T_i(P, H)$ as the set of possible states of mind agent $i$ could be in if he considered $H$ possible. Consider the following three properties of generalized hypothesis transformations:

T0. $T_i(P, H)$ is a singleton.

T1$'$. $P' \cap H \neq \emptyset$ for all $P' \in T_i(P, H)$.

T2$'$. If $P \cap H \neq \emptyset$, then $T_i(P, H) = \{P\}$.

T1$'$ and T2$'$ are the obvious generalizations of T1 and T2. If we assume T0, then we are back in the setting of hypothesis transformations and extended information structures.

In a generalized information structure, we can define the binary operator $K_i$ as follows:

$$K_i^H(E) = \cup\{P \in \Pi_i : \cup T_i(P, H) \subseteq E\}.$$

Clearly this definition generalizes (2).

We now have the following generalization of Theorem 2.2.

**Theorem 4.5:** *Let* $\mathcal{T}$ *be a (possibly empty) subset of* $\{T0, T1', T2'\}$*, let* $\mathcal{K}$ *be the corresponding subset of* $\{K3, K5, K6\}$*, and let* $(\Omega, \Pi_1, \ldots, \Pi_n)$ *be an information structure. All generalized hypothesis transformations* $T_1, \ldots, T_n$ *on this information structure satisfying the assumptions in* $\mathcal{T}$ *define hypothetical knowledge operators that satisfy K1, K2, K4, K7, and the axioms in* $\mathcal{K}$*. Moreover, if* $K : (2^\Omega - \emptyset) \times 2^\Omega \rightarrow 2^\Omega$ *is a binary operator satisfying K1, K2, K4, K7, and the axioms in* $\mathcal{K}$ *(for agent* $i$*), then there is a generalized hypothesis transformation* $T_i$ *on* $\Pi_i$ *satisfying* $\mathcal{T}$ *such that* $K_i^H(E) = \cup\{P \in \Pi_i : \cup T_i(P, H) \subseteq E\}$.

As a corollary to this result, it follows that there is a direct correspondence between counterfactual information structures and generalized information structures such that $K_i = \overline{\overline{K}}_i$. More precisely, we have the following result.

**Theorem 4.6:** *Let $\mathcal{S}$ be a (possibly empty) subset of {R1,R3,R45} and let $\mathcal{T}$ be the corresponding subset of {T0,T1',T3'}. If $C = (\Omega, \Pi_1, \ldots, \Pi_n, f_1, \ldots, f_n)$ is a counterfactual information structure such that the selection functions $f_i$ satisfy R2 and the properties in $\mathcal{S}$, then there exist generalized hypothesis transformations $T_1, \ldots, T_n$ such that $\overline{\overline{K}}_i^H(E) = K_i^H(E)$ (where $\overline{\overline{K}}_i$ is defined in the counterfactual information structure $C$ and $K_i$ is defined in the generalized information structure $(\Omega, \Pi_1, \ldots, \Pi_n, T_1, \ldots, T_n)$). Similarly, if $G = (\Omega, \Pi_1, \ldots, \Pi_n, T_1, \ldots, T_n)$ is a generalized information structure such that the generalized hypothesis transformations satisfy the properties in $\mathcal{T}$, then there exist selection functions $f_1, \ldots, f_n$ satisfying R2 and the properties in $\mathcal{S}$ such that $\overline{\overline{K}}_i^H(E) = K_i^H(E)$.*

**Proof:** For the first part, note that by Theorem 4.4, the $\overline{\overline{K}}_i$ operators in $C$ satisfy K1, K2, K4, K7 and the subset of {K5,K6,K3} corresponding to {R1,R3,R45}. Thus, it follows from Theorem 4.5 that there exist generalized transformation operators $T_i$ such that $\overline{\overline{K}}_i^H(E) = K_i^H(E)$. (In fact, we can define $T_i(P, H) = \{\Pi(\omega') : \omega' \in f_i(\omega, L_i(H))$, for some $\omega \in P\}$.) The proof of the second half is similar. ▮

# 5 Discussion

How reasonable is Samet's framework? The translation of it into conditional logic suggests that it suffers from three potential problems: (1) unreasonable assumptions, (2) lack of expressive power, and (3) missing axioms. I discus the first two issues here; for a discussion of the third issue, see the full paper.

## 5.1 A Closer Look at R4 and R5

Using our translation(s), all of Samet's axioms follow from minimal assumptions in the standard framework for modeling counterfactuals except for K3 (and K1 if we use $\overline{K}_i^H$). To get these properties, we need both R4 and R5, whichever translation we use. As the following examples show, neither is a reasonable assumption in general, and hence neither is K3.

**Example 5.1:** Consider again Example 2.3. As we observed, there seems to be no appropriate definition for $T_1$ in this case, since $T_1$ is required to be a function. On the other hand, taking $f((\mathit{red}, \mathit{off}), \mathit{ON}) = \{(\mathit{red}, \mathit{on})\}$ and $f((\mathit{blue}, \mathit{off}), \mathit{ON}) = \{(\mathit{blue}, \mathit{on})\}$ seems to capture the story, but it does not satisfy R5.

This shows why we may not want an agent to know his selection function, in the sense that it is the same at all worlds that he consider possible. In general, we may want the selection function to depend on features of the actual world (and yet still be subjective). With this selection function, we have that $(\mathit{red}, \mathit{off}) \in \mathit{OFF} >_1 \mathit{RED}$ while $(\mathit{blue}, \mathit{off}) \notin \mathit{OFF} >_1 \mathit{RED}$, although the agent cannot distinguish $(\mathit{red}, \mathit{off})$ from $(\mathit{blue}, \mathit{off})$. Translated to English, this says that, according to agent 1, the question of whether or not he would see a red door if the light were on depends on whether the door is actually red. As a result, $\overline{K}_1^{ON}(RED) \neq K_1(\overline{K}_1^{ON}(RED))$, violating K1.

We could capture this situation by means of a *generalized* hypothesis transformation $T$, that maps a cell to a union of cells, rather than a unique cell. For example, we could have $T(OFF, ON) = ON$ (note that $T = \overline{f}$). The intuition here is that there may be more than one cell closest to *OFF* where *ON* is true. ▮

Example 5.1 shows the problem with R5; the next example shows the problem with R4.

**Example 5.2:** Suppose we have a simple game where player 1 can either go left or right. If he goes left, he gets a payoff of 3; if he goes right, he gets a payoff of either 1 or 4, depending on whether player 2 goes left or right. For simplicity, suppose that player 2's payoff is 1 no matter what she does. Thus, there are three worlds, call them $L$, $RL$, and $RR$, with the obvious interpretation. Let $R = \{RL, RR\}$ be the event where player 1 goes right. Agent 1 has two information sets: $\{L\}$ and $R$. What should $f(L, R)$ be? It seems reasonable in this case to take $f_1(L, R) = R$: if player 1 had gone right, he has no idea whether player 2 would have gone left or right. But with this choice, R4 does not hold. Not surprisingly, K3 does not hold either: $\neg\overline{K}_1^R(\{RL\}) \neq \overline{K}_1^R(\neg K_1(\{RL\}))$ (and similarly if we replace $\overline{K}$ by $\overline{\overline{K}}$).

In this case, the weaker version of R4 discussed in Footnote 3, which only requires $f_1(L, R)$ to be a subset of some cell, does hold. However, a variant of this example shows that even the weaker version of R4 is not so reasonable. Suppose that, again, if he goes left, player 1 gets a payoff of 3, but if he goes right, he must choose a number, either 0 or 1. Then player 2 chooses a number. If both players choose the same number, player 1 gets a payoff of 4, otherwise he gets a payoff of 1. Again, suppose that player 2's payoff is 1 no matter what. We can now consider a model with five worlds, one corresponding to each play of this game. We can take $L$ to be the world where player 1 plays left, and $R$ to be the event consisting of the remaining 4 worlds, where player 1 plays right. If we assume that player 1 knows his action, then R splits into two information sets, one where player 1 chooses 0 and one where he chooses 1. Nevertheless, it still seems reasonable to take $f_1(L, R) = R$. Assuming that player 1 randomizes after going right, at $L$, why should player 1 know what action he would have chosen if he had gone right?
▮

These examples reinforce the case against K3. So where does this leave Samet's results on backwards induction? In fact, these results hold even without K3. I briefly review Samet's framework in order to explain this point. Samet does not take strategies as primitive; rather, he defines strategies in terms of counterfactuals. To define strategies, he makes use of statements of the form "If I were at information set $I$ in the game tree, I would perform action $a$". Let $H_I$ be the statement "I am at information set $I$" and let $E_a$ be the statement "I perform action $a$".[5] Assumption K3 is required to guarantee that for each information set $I$, there is a *unique*

[5] Actually, Samet considers statements of the form "if I were at node $n$, I would perform action $a$". For games of perfect information, the information sets are nodes. In games of imperfect information, we want the counterfactual statement to involve information sets, not nodes.

action $a$ such that $K^{H_I}(E_a)$ holds. If Samet had used generalized information sets (where the generalized hypothesis transformations satisfy T1$'$ and T2$'$, but not necessarily T0), rather than extended information sets, this approach would have led to *nondeterministic* strategies, where at each information set there is more than one possible action that the agent might perform. (If we further put a probability measure on the information sets in $T_i(P, H)$, this approach would result in a behavior strategy.)

Samet uses his notion of strategy (and other counterfactual reasoning) to show that in a *nondegenerate* game of perfect information (that is, a game with different payoffs at each of the terminal nodes), *a common hypothesis of node rationality* (see [Samet 1996] for the formal definition of this notion) implies that the players play the backwards induction solution. His proof would go through with essentially no change if he had used generalized information structures (thus dropping K3). A common hypothesis of node rationality would still suffice to give us the backwards induction solution, even if we start with nondeterministic strategies.[6]

## 5.2 Restricted expressive power

The translations show that Samet's hypothetical knowledge operators correspond to rather restricted expressions of conditional logic, those of the form $L_i(H) >_i K_i(H)$ or $K_i(L_i(H) >_i K_i(H))$, depending on the translation. While such expressions may suffice to deal with the particular situations considered by Samet (but see the concerns expressed in the previous subsection), they do not suffice for general game-theoretic reasoning. For example, suppose in a game of imperfect information, player 1 cannot distinguish nodes $n_1$ and $n_2$ (that is, they are in the same cell of player 1's partition) and considers it possible that both will be reached, but player 2 can distinguish them. Player 1 might well want to make a statement such as "I know that if we were to reach node $n_1$ ($H$), then player 2 would play action $a$". Such a statement corresponds to an event of the form $K_1(H >_1 E)$, but does not correspond to an event of the form $K_1^{H'}(E')$, for any choice of $H'$ or $E'$. It is not necessarily true that if player 1 considers it possible that node $n_1$ is reached, then player 2 would play $a$, since at node $n_2$ player 2 might not play $a$.

Once we recognize the need for such statements, we are forced to go to the more expressive formalism of conditional logic combined with epistemic logic, rather than Samet's formalism.

# 6 Conclusion

My goal here was relatively modest: simply to show that Samet's hypothetical knowledge operators could be captured using the standard models of conditional logic and epistemic logic. I hope that the reader is now convinced that this can be done. By using more standard means to capture hypothetical knowledge, we can see the potential problems with K3 (and its

[6] Samet [personal communication, 1998] observes that in fact the main role of K3 is to simplify the formulation and proofs of his results.

connection to well-known issues in philosophical logic, going back to the discussion of whether the selection function should always return a unique closest world). Moreover, we have access to a richer language which, as I tried to suggest in Section 5.2, may prove useful in analyzing games of imperfect information.

**Acknowledgments** I'd like to thank Nir Friedman and Dov Samet for their valuable comments on and criticism of an earlier version of this paper, and the two anonymous referees for their comments.

## References

Aumann, R. J. (1995). Backwards induction and common knowledge of rationality. *Games and Economic Behavior 8*, 6–19.

Binmore, K. (1996). Rationality and backward induction. Unpublished manuscript, available at http://ada.econ.ucl.ac.uk/papers.htm. This is a revised version of "Rationality in the Centipede", which appears in *Proc. 4th Conference on Theoretical Aspects of Reasoning About Knowledge* (R. Fagin, ed.), 1994, pp. 150–159.

Halpern, J. Y. (1998). Semantic completeness for epistemic and conditional logic. In *Fifth International Conference on AI and Mathematics*. Proceedings available at http://rutcor.rutgers.edu/ amai.

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

Samet, D. (1994). The logic of hypothetical knowledge. Unpublished manuscript.

Samet, D. (1996). Hypothetical knowledge and games with perfect information. *Games and Economic Behavior 17*, 230–251.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory*, Number 2 in American Philosophical Quarterly monograph series. Blackwell, Oxford. Also appears in *Ifs* (Ed. W. L. Harper, R. C. Stalnaker and G. Pearce), Reidel, Dordrecht, Netherlands, 1981.

Stalnaker, R. C. (1992). Notes on conditional semantics. In Y. Moses (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fourth Conference*, pp. 316–328. San Francisco, Calif.: Morgan Kaufmann.

Stalnaker, R. C. and R. Thomason (1970). A semantical analysis of conditional logic. *Theoria 36*, 246–281.