Philosophical Studies (2006) 128:169–199
DOI 10.1007/s11098-005-4062-y

ROBERT STALNAKER

ON LOGICS OF KNOWLEDGE AND BELIEF

1. INTRODUCTION

Formal epistemology, or at least the approach to formal epistemology that develops a logic and formal semantics of knowledge and belief in the possible worlds framework, began with Jaakko Hintikka's book *Knowledge and Belief*, published in 1962. Hintikka's project sparked some discussion of issues about iterated knowledge (does knowing imply knowing that one knows?) and about "knowing who", and quantifying into knowledge attributions. Much later, this kind of theory was taken up and applied by theoretical computer scientists and game theorists.[1] The formal semantic project gained new interest when it was seen that it could be applied to contexts with multiple knowers, and used to clarify the relation between epistemic and other modal concepts.

Edmund Gettier's classic refutation of the Justified True Belief analysis of knowledge (Gettier, 1963) was published at about the same time as Hintikka's book, and it immediately spawned an epistemological industry – a project of attempting to revise the refuted analysis by adding further conditions to meet the counterexamples. Revised analyses were met with further counterexamples, followed by further refinements. This kind of project flourished for some years, but eventually became an internally driven game that was thought to have lost contact with the fundamental epistemological questions that originally motivated it. This way of approaching epistemological questions now seems hopelessly out of date, but I think there may still be some insights to be gained by looking back, if not at the details of the analyses, at least at some of the general strategies of analysis that were deployed.

There was little contact between these two very different epistemological projects. The first had little to say about substantive questions about the relation between knowledge, belief, and justification or epistemic entitlement, or about traditional epistemological issues, such as skepticism. The second project ignored questions about the abstract structure of epistemic and doxastic states. But I think some of the abstract questions about the logic of knowledge connect with traditional questions in epistemology, and with the issues that motivated the attempt to find a definition of knowledge. The formal semantic framework provides the resources to construct models that may help to clarify the abstract relationship between the concept of knowledge and some of the other concepts (belief and belief revision, causation and counterfactuals) that were involved in the post-Gettier project of defining knowledge. And some of the examples that were originally used in the post-Gettier literature to refute a proposed analysis can be used in a different way in the context of formal semantic theories: to bring out contrasting features of some alternative conceptions of knowledge, conceptions that may not provide plausible analyses of knowledge generally, but that may provide interesting models of knowledge that are appropriate for particular applications, and that may illuminate, in an idealized way, one or another of the dimensions of the complex epistemological terrain.

My aim in this paper will be to bring out some of the connections between issues that arise in the development and application of formal semantics for knowledge and belief and more traditional substantive issues in epistemology. The paper will be programmatic, pointing to some highly idealized theoretical models, some alternative assumptions that might be made about the logic and semantics of knowledge, and some of the ways in which they might connect with traditional issues in epistemology, and with applications of the concept of knowledge. I will bring together and review some old results, and make some suggestions about possible future developments. After a brief sketch of Hintikka's basic logic of knowledge, I will discuss, in Section 2, the S5 epistemic

models that were developed and applied by theoretical computer scientists and game theorists, models that, I will argue, conflate knowledge and belief. In Section 3, I will discuss a basic theory that distinguishes knowledge from belief and that remains relatively noncommittal about substantive questions about knowledge, but that provides a definition of belief in terms of knowledge. This theory validates a logic of knowledge, S4.2, that is stronger than S4, but weaker than S5. In the remaining four sections, I will consider some alternative ways of adding constraints on the relation between knowledge and belief that go beyond the basic theory: in Section 4, I will consider the S5 partition models as a special case of the basic theory; in Section 5, I will discuss the upper and lower bounds to an extension of the semantics of belief to a semantics for knowledge; in Section 6, I will discuss a version of the defeasibility analysis of knowledge, and in Section 7, a simplified version of a causal theory.

The basic idea that Hintikka developed, and that has since become familiar, was to treat knowledge as a modal operator with a semantics that parallels the possible worlds semantics for necessity. Just as necessity is truth in all possible worlds, so knowledge is truth in all *epistemically* possible worlds. The assumption is that to have knowledge is to have a capacity to locate the actual world in logical space, to exclude certain possibilities from the candidates for actuality. The epistemic possibilities are those that remain after the exclusion, those that the knower cannot distinguish from actuality. To represent knowledge in this way is of course not to provide any kind of reductive analysis of knowledge, since the abstract theory gives no substantive account of the criteria for determining epistemic possibility. The epistemic possibilities are defined by a binary accessibility relation between possible worlds that is a primitive component of an epistemic model. (Where x and y are possible worlds, and 'R' is the accessibility relation, 'xRy' says that y is epistemically possible for the agent in world x). The idea was to give a precise representation of the structure of an epistemic state that was more or less neutral about more substantive questions about what

constitutes knowledge, but that sharpened questions about the logic of knowledge. This form of representation was, however, far from innocent, since it required, from the start, an extreme idealization: Even in its most neutral form, the framework required the assumption that knowers know all logical truths, and all of the consequences of their knowledge, since no matter how the epistemically possible worlds are selected, all logical truths will be true in all of them, and for any set of propositions true in all of them, all of their logical consequences will also be true in all of them. There are different ways of understanding the character of this idealization: on the one hand, one might say that the concept of knowledge that is being modeling is knowledge in the ordinary sense, but that the theory is intended to apply only to idealized knowers – those with superhuman logical capacities. Alternatively, one might say that the theory is intended to model an idealized sense of knowledge – the information that is implicit in one's knowledge – that literally applies to ordinary knowers. However the idealization is explained, there remain the questions whether it is fruitful to develop a theory that requires this kind of deviation from reality, and if so why.[2] But I think these questions are best answered by looking at the details of the way such theories have been, and can be developed.

The most basic task in developing a semantics for knowledge in the possible worlds framework is to decide on the properties of the epistemic accessibility relation. It is clear that the relation should be reflexive, which is necessary to validate the principle that knowledge implies truth, an assumption that is just about the only principle of a logic of knowledge that is uncontroversial. Hintikka argued that we should also assume that the relation is transitive, validating the much more controversial principle that knowing implies knowing that one knows. Knowing and knowing that one knows are, Hintikka claimed, "virtually equivalent." Hintikka's reasons for this conclusion were not completely clear. He did not want to base it on a capacity for introspection: he emphasized that his reasons were logical rather than psychological. His proof of the KK principle rests on the

following principle: If $\{K\phi, \sim K\sim\psi\}$ is consistent, then $\{K\phi, \psi\}$ is consistent, and it is clear that if one grants this principle, the KK principle immediately follows.[3] The reason for accepting this principle seems to be something like this: Knowledge requires conclusive reasons for belief, reasons that would not be defeated by any information compatible with what is known. So if one knows that ϕ while ψ is compatible with what one knows, then the truth of ψ could not defeat one's claim to know that ϕ. This argument, and other considerations for and against the KK principle deserve more careful scrutiny. There is a tangle of important and interesting issues underlying the question whether one should accept the KK principle, and the corresponding semantics, and some challenging arguments that need to be answered if one does.[4] I think the principle can be defended (in the context of the idealizations we are making), but I will not address this issue here, provisionally following Hintikka in accepting the KK principle, and a semantics that validates it.

The S4 principles (Knowledge implies truth, and knowing implies knowing that one knows) were as far as Hintikka was willing to go. He unequivocally rejects the characteristic S5 principle that if one lacks knowledge, then one knows that one lacks it. ("unless you happen to be as sagacious as Socrates"[5]), and here his reasons seem to be clear and decisive:[6]

> The consequences of this principle, however, are obviously wrong. By its means (together with certain intuitively acceptable principles) we could, for example, show that the following sentence is self sustaining:
>
> (13) $p \supset K_a P_a p$ [In Hintikka's notation, 'P_a' is the dual of the knowledge operator, 'K_a': '$\sim K_a\sim$'. I will use 'M' for $\sim K\sim$]

The reason that (13) is clearly unacceptable, as Hintikka goes on to say, is that it implies that one could come to know by reflection alone, of any truth, that it was compatible with one's knowledge. But it seems that a consistent knower might believe, and be justified in believing, that she knew something that was in fact false. That is, it might be, for some proposition

ϕ that $\sim\phi$, and $BK\phi$. In such a case, if the subject's beliefs are consistent, then she does not believe, and so does not know, that $\sim\phi$ is compatible with her knowledge. That is, $\sim K\sim K\phi$, along with $\sim\phi$, will be true, falsifying (13).

2. PARTITION MODELS

Despite Hintikka's apparently decisive argument against the S5 principle, later theorists applying epistemic logic and semantics, both in theories of distributive computer systems and in game theory assumed that S5 was the right logic for (an idealized concept of) knowledge, and they developed semantic models that seem to support that decision. But while such models, properly interpreted, have their place, I will argue that they have conflated knowledge and belief in a way that has led to some conceptual confusion, and that they have abstracted away from some interesting problems within their intended domains of application that more general models might help to clarify. But before getting to this issue, let me first take note of another way that more recent theorists have modified, or generalized, Hintikka's original theory.

Hintikka's early models were models of the knowledge of a single knower, but much of the later interest in formal epistemic models derives from a concern with situations in which there are multiple knowers who may know or be ignorant about the knowledge and ignorance of the others. While Hintikka's early work did not give explicit attention to the interaction of different knowers, the potential to do so is implicit in his theory. Both the logic and the semantics of the knowledge of a single knower generalize in a straightforward way to a model for multiple knowers. One needs only a separate knowledge operator for each knower, and in the semantics, a separate relation of epistemic accessibility for each knower that interprets the operator. One can also introduce, for any group of knowers, an operator for the *common* knowledge shared by the member of the group, where a group has common knowledge that ϕ if and only if all

know that ϕ, all know that all know that ϕ, all know that all know that all know, etc. all the way up. The semantics for the common knowledge operator is interpreted in terms of an accessibility relation that is definable in terms of the accessibility relations for the individual knowers: the common-knowledge accessibility relation for a group G is the transitive closure of the set of epistemic accessibility relations for the members of that group.[7] If R^G is this relation, then the knowers who are members of G have common knowledge that ϕ (in possible world x) if ϕ is true in all possible worlds that are R^G related to world x. The generalization to multiple knowers and to common knowledge, works the same way, whatever assumptions one makes about the accessibility relation, and one can define notions of common belief in an exactly analogous way. The properties of the accessibility relations for common knowledge and common belief will derive from the properties of the individual accessibility relations, but they won't necessarily be the same as the properties of the individual accessibility relations. (Though if the logic of knowledge is S4 or S5, then the logic of common knowledge will also be S4 or S5, respectively).

Theoretical computer scientists have used the logic and semantics for knowledge to give abstract descriptions of distributed computer systems (such as office networks or email systems) that represent the distribution and flow of information among the components of the system. For the purpose of understanding how such systems work and how to design protocols that permit them to accomplish the purposes for which they are designed, it is useful to think of them as communities of interacting rational agents who use what information they have about the system as a whole to serve their own interests, or to play their part in a joint project. And it is useful in turn for those interested in understanding the epistemic states of rational agents to think of them in terms of the kind of simplified models that theoretical computer scientists have constructed.

A distributed system consists of a set of interconnected components, each capable of being in a range of local states.

The way the components are connected, and the rules by which the whole system works, constrain the configurations of states of the individual components that are possible. One might specify such a system by positing a set of n components and possible local states for each. One might also include a component labeled "nature" whose local states represent information from outside the system proper. *Global* states will be n-tuples of local states, one for each component, and the model will also specify the set of global states that are *admissible*. Admissible global states are those that are compatible with the rules governing the way the components of the system interact. The admissible global states are the possible worlds of the model. This kind of specification will determine, for each local state that any component might be in, a set of global states (possible worlds) that are compatible with the component being in that local state. This set will be the set of epistemically possible worlds that determines what the component in that state knows about the system as a whole.[8] Specifically, if 'a' and 'b' denote admissible global states, and 'a_i'and 'b_i' denote the ith elements of a and b, respectively (the local states of component i), then global world-state b is epistemically accessible (for i) to global world-state a if and only if $a_i = b_i$. So, applying the standard semantic rule for the knowledge operator, component (or knower) i will know that ϕ, in possible world a, if and only if ϕ is true in all possible worlds in which i has the same local state that it has in world-state a. One knows that ϕ if one's local state carries the information that ϕ.[9]

Now it is obvious that this epistemic accessibility relation is an equivalence relation, and so the logic for knowledge in a model of this kind is S5. Each of the epistemic accessibility relations partitions the space of possible worlds, and the cross-cutting partitions give rise to a simple and elegant model of common knowledge, also with an S5 logic. Game theorists independently developed this kind of partition model of knowledge and have used such models to bring out the consequences of assumptions about common knowledge. For example, it can be shown that, in certain games, players will always

make certain strategy choices when they have common knowledge that all players are rational. But as we have seen, Hintikka gave reasons for rejecting the S5 logic for knowledge, and the reasons seemed to be decisive. It seems clear that a consistent and epistemically responsible agent might take herself to know that ϕ in a situation in which ϕ was in fact false. Because knowledge implies truth, it would be false, in such a case, that the agent knew that ϕ, but the agent could not know that she did not know that ϕ without having inconsistent beliefs. If such a case is possible, then there will be counterexamples to the S5 principle, $\sim K\phi \rightarrow K\sim K\phi$. That is, the S5 principles require that rational agents be immune to error. It is hard to see how any theory that abstracts away from the possibility of error could be relevant to epistemology, an enterprise that begins with skeptical arguments using scenarios in which agents are systematically mistaken and that seeks to explain the relation between knowledge and belief, presupposing that these notions do not coincide.

Different theorists have different purposes, and it is not immediately obvious that the models of knowledge that are appropriate to the concerns of theoretical computer scientists and game theorists need be relevant to issues in epistemology. But I think that the possibility of error, and the differences between knowledge and belief are relevant to the intended domains of application of those models, and that some of the puzzles and problems that characterize epistemology are reflected in problems that may arise in applying those theories.

As we all know too well, computer systems sometimes break down or fail to behave as they were designed to behave. In such cases, the components of a distributed system will be subject to something analogous to error and illusion. Just as the epistemologist wants to explain how and when an agent knows some things even when he is in error about others, and is interested in methods of detecting and avoiding error, so the theoretical computer scientist is interested in the way that the components of a system can avoid and detect faults, and can continue to function appropriately even when conditions are not completely normal. To clarify such

problems, it is useful to distinguish knowledge from something like belief.

The game theorist, or any theorist concerned with rational action, has a special reason to take account of the possibility of false belief, even under the idealizing assumption that in the actual course of events, everyone's beliefs are correct. The reason is that decision theorists and game theorists need to be concerned with causal or counterfactual possibilities, and to distinguish them from epistemic possibilities. When I deliberate, or when I reason about why it is rational to do what I know that I am going to do, I need to consider possible situations in which I make alternative choices. I know, for example, that it would be irrational to cooperate in a one-shot prisoners' dilemma because I know that in the counterfactual situation in which I cooperate, my payoff is less than it would be if I defected. And while I have the capacity to influence my payoff (negatively) by making this alternative choice, I could not, by making this choice, influence your prior beliefs about what I will do; that is, your prior beliefs will be the same, in the counterfactual situation in which I make the alternative choice, as they are in the actual situation Since you take yourself (correctly, in the actual situation) to know that I am rational, and so that I will not cooperate, you therefore also take yourself to know, in the counterfactual situation I am considering, that I am rational, and so will not cooperate. But in that counterfactual situation, you are wrong – you have a false belief that you take to be knowledge. There has been a certain amount of confusion in the literature about the relation between counterfactual and epistemic possibilities, and this confusion is fed, in part, by a failure to make room in the theory for false belief.[10]

Even in a context in which one abstracts away from error, it is important to be clear about the nature of the idealization, and there are different ways of understanding it that are sometimes confused. But before considering the alternative ways of making the S5 idealization, let me develop the contrast between knowledge and belief, and the relation between them, in a more general setting.

3. BELIEF AND KNOWLEDGE

Set aside the S5 partition models for the moment, and consider, from a more neutral perspective, the logical properties of belief, and the relation between belief and knowledge. It seems reasonable to assume, at least in the kind of idealized context we are in, that agents have introspective access to their beliefs: if they believe that ϕ, then they know that they do, and if they do not, then they know that they do not. (The S5, "negative introspection" principle, $\sim K\phi \rightarrow K\sim K\phi$, was problematic for knowledge because it is in tension with the fact that knowledge implies truth, but the corresponding principle for belief does not face this problem.) It also seems reasonable to assume that knowledge implies belief. Given the fact that our idealized believers are logically omniscient, we can assume, in addition, that their beliefs will be consistent. Finally, to capture the fact that our intended concept of belief is a strong one – subjective certainty – we assume that believing implies believing that one knows. So our logic of knowledge and belief should include the following principles in addition to those of the logic S4:

(PI)	$\vdash B\phi \rightarrow \mathrm{KB}\phi$	Positive introspection
(NI)	$\vdash \sim B\phi \rightarrow K\sim B\phi$	Negative introspection
(KB)	$\vdash K\phi \rightarrow B\phi$	Knowledge implies belief
(CB)	$\vdash B\phi \rightarrow \sim B\sim\phi$	Consistency of belief
(SB)	$\vdash B\phi \rightarrow \mathrm{BK}\phi$	Strong belief

The resulting combined logic for knowledge and belief yields a pure belief logic, KD45, which is validated by a doxastic accessibility relation that is serial, transitive and euclidean.[11] More interestingly, one can prove the following equivalence theorem: $B\phi \leftrightarrow \mathrm{MK}\phi$ (using 'M' as the epistemic possibility operator, '$\sim K\sim$'). This equivalence permits a more economical formulation of the combined belief-knowledge logic in which the belief operator is defined in terms of the knowledge operator. If we substitute 'MK' for 'B' in our principle (CB), we get $\mathrm{MK}\phi \rightarrow \mathrm{KM}\phi$, which, if added

to S4 yields the logic of knowledge, S4.2. All of the other principles listed above (with 'MK' substituted for '*B*') are theorems of S4.2, so this logic of knowledge by itself yields a combined logic of knowledge and belief with the appropriate properties.[12]

The assumptions that are sufficient to show the equivalence of belief with the epistemic possibility of knowledge (one believes that ϕ, in the strong sense, if and only if it is compatible with one's knowledge that one knows that ϕ) might also be made for a concept of *justified* belief, although the corresponding assumptions will be more controversial. Suppose (1) one assumes that justified belief is a necessary condition for knowledge, and (2) one adopts an *internalist* conception of justification that supports the positive and negative introspection conditions (if one has justified belief that ϕ, one knows that one does, and if one does not, one knows that one does not), and (3) one assumes that since the relevant concept of belief is a strong one, one is justified in believing that ϕ if and only if one is justified in believing that one knows that ϕ. Given these assumptions, justified belief will also coincide with the epistemic possibility that one knows, and so belief and justified belief will coincide. The upshot is that for an internalist, a divergence between belief (in the strong sense) and justified belief would be a kind of internal inconsistency. If one is not fully justified in believing ϕ, one knows this, and so one knows that a necessary condition for knowledge that ϕ is lacking. But if one believes that ϕ, in the strong sense, then one believes that one knows it. So one both knows that one lacks knowledge that ϕ, and believes that one has knowledge that ϕ.

The usual constraint on the accessibility relation that validates S4.2 is the following convergence principle (added to the transitivity and reflexivity conditions): if xRy and xRz, then there is a w such that yRw and zRw. But S4.2 is also sound and complete relative to the following stronger convergence principle: for all x, there is a y such that for all z, if xRz, then zRy. The weak convergence principle (added to reflexivity and transitivity) implies that for any *finite* set of worlds accessible to x, there is a single world accessible with

respect to all of them. The strong convergence principle implies that there is a world that is accessible to *all* worlds that are accessible to x. The semantics for our logic of knowledge requires the stronger convergence principle.[13]

Just as, within the logic, one can define belief in terms of knowledge, so within the semantics, one can define a doxastic accessibility relation for the derived belief operator in terms of the epistemic accessibility relation. If 'R' denotes the epistemic accessibility relation and 'D' denotes the doxastic relation, then the definition is as follows: $xDy =_{\text{df}} (z)(xRz \rightarrow zRy)$. Assuming that R is transitive, reflexive and strongly convergent, it can be shown that D will be serial, transitive and euclidean – the constraints on the accessibility relation that characterize the logic KD45.

One can also define, in terms of D, and so in terms of R, a third binary relation on possible worlds that is relevant to describing the epistemic situation of our ideal knower: Say that two possible worlds x and y are *epistemically indistinguishable* to an agent (xEy) if and only if she has exactly the same beliefs in world x as she has in world y. That is, $xEy =_{\text{df}} (z)(xDz \leftrightarrow yDz)$. E is obviously an equivalence relation, and so any modal operator interpreted in the usual way in terms of E would be an S5 operator But while this relation is definable in the semantics in terms of the epistemic accessibility relation, we cannot define, in the object language with just the knowledge operator, a modal operator whose semantics is given by this accessibility relation.

So the picture that our semantic theory paints is something like this: For any given knower i and possible world x, there is, first, a set of possible worlds that are subjectively indistinguishable from x, to i (those worlds that are E-related to x); second, there is a subset of that set that includes just the possible worlds compatible with what i *knows* in x (those worlds that are R-related to x); third, there is a subset of that set that includes just the possible worlds that are compatible with what i *believes* in x (those worlds that are D-related to x). The world x itself will necessarily be a member of the outer set and of the R-subset, but will not necessarily be a member

of the inner D-subset. But if x is itself a member of the inner D-set (if world x is itself compatible with what i believes in x), then the D-set will coincide with the R-set.

Here is one way of seeing this more general theory as a generalization of the distributive systems models, in which possible world-states are sequences of local states: one might allow *all* sequences of local states (one for each agent) to count as possible world-states, but specify, for each agent, a subset of them that are *normal* – the set in which the way that agent interacts with the system as a whole conforms to the constraints that the system conforms to when it is functioning as it is supposed to function. In such models, two worlds, x and y, will be subjectively indistinguishable, for agent i (xE_iy), whenever $x_i = y_i$ (so the relation that was the epistemic accessibility relation in the unreconstructed S5 distributed systems model is the subjective indistinguishability relation in the more general models). Two worlds are related by the doxastic accessibility relation (xD_iy) if and only if $x_i = y_i$, *and in addition, y is a normal world, with respect to agent i.*[14] This will impose the right structure on the D and E relations, and while it imposes some constraints on the epistemic accessibility relation, it leaves it underdetermined. We might ask whether R can be defined in a plausible way in terms of the components of the model we have specified, or whether one might add some independently motivated components to the definition of a model that would permit an appropriate definition of R. This question is a kind of analogue of the question asked in the more traditional epistemological enterprise – the project of giving a definition of knowledge in terms of belief, truth, justification, and whatever other normative and causal concepts might be thought to be relevant. Transposed into the model theoretic framework, the traditional problem of adding to true belief further conditions that together are necessary and sufficient for knowledge is the problem of extending the doxastic accessibility relation to a reflexive relation that is the right relation (at least in the idealized context) for the interpretation of a knowledge operator. In the remainder of this paper, I will

consider several ways that this might be done, and at the logics of knowledge that they validate.

4. PARTITION MODELS AND THE BASIC THEORY

One extreme way of defining the epistemic accessibility relation in terms of the resources of our models is to identify it with the relation of subjective indistinguishability, and this is one way that the S5 partition models have implicitly been interpreted. If one simply assumes that the epistemic accessibility relation is an equivalence relation, this will suffice for a collapse of our three relations into one. Subjective indistinguishability, knowledge, and belief will all coincide. This move imposes a substantive condition on knowledge, and so on belief, when it is understood in the strong sense as belief that one knows, a condition that is appropriate for the skeptic who thinks that we are in a position to have genuine knowledge only about our own internal states – states about which we cannot coherently be mistaken. On this conception of knowledge, one can have a false belief (in the strong sense) only if one is internally inconsistent, and so this conception implies a bullet-biting response to the kind of argument that Hintikka gave against the S5 logic for knowledge. Hintikka's argument was roughly this: S5 validates the principle that any proposition that is in fact true, is known by any agent to be compatible with his knowledge, and this is obviously wrong: The response suggested by the conception of knowledge that identifies knowledge with subjective indistinguishability is that if we assume that all we can know is how things seem to us, and also assume that we are infallible judges of the way things seem to us, then it will be reasonable to conclude that we are in a position to know, of anything that is in fact false, that we do not know it.

There is a less radical way to reconcile our basic theory of knowledge and belief with the S5 logic and the partition models. Rather than making more restrictive assumptions about the concept of knowledge, or about the basic structure of the model, one may simply restrict the intended domain of application of the theory to cases in which the agent in

question has, in fact, only true beliefs. On this way of understanding the S5 models, the model theory does not further restrict the relations between the three accessibility relations, but instead assumes that the *actual* world of the model is a member of the inner *D*-set.[15] This move does not provide us with a way to define the epistemic accessibility relation in terms of the other resources of the model; but what it does is to stipulate that the actual world of the model is one for which the epistemic accessibility relation is determined by the other components. (That is, the set of worlds *y* that are epistemically accessible to the actual world is determined) Since the assumptions of the general theory imply that all worlds outside the *D*-sets are epistemically inaccessible to worlds within the *D*-sets, and that all worlds within a given *D*-set are epistemically accessible to each other, the assumption that the actual world of the model is in a *D*-set will determine the *R*-set for the actual world, and will validate the logic S5.

So long as the object language that is being interpreted contains just one modal operator, an operator representing the knowledge of a single agent, the underdetermination of epistemic accessibility will not be reflected in the truth values in a model of any expressible proposition. Since all possible worlds outside of any *D*-set will be invisible to worlds within it, one could drop them from the model (taking the set of all possible worlds to be those *R*-related to the actual world) without affecting the truth values (at the actual world) of any sentence. This generated submodel will be a simple S5 model, with a universal accessibility relation. But as soon as one enriches the language with other modal and epistemic operators, the situation changes. In the theory with two or more agents, even if one assumes that all agents have only true beliefs, the full S5 logic will not be preserved. The idealizing assumption will imply that Alice's beliefs coincide with her knowledge (in the actual world), and that Bob's do as well, but it will not follow that Bob knows (in the actual world) that Alice's beliefs coincide with her knowledge. To validate the full S5 logic, in the multiple agent theory, we need to assume that it is not just true, but common knowledge that everyone has only true

beliefs. This stronger idealization is needed to reconcile the partition models, used in both game theory and in distributed systems theory, with the general theory that allows for a distinction between knowledge and belief. But even in a context in which one makes the strong assumption that it is common knowledge that no one is in error about anything, the possible divergence of knowledge and belief, and the failure of the S5 principles to be *necessarily* true will show itself when the language of knowledge and common knowledge is enriched with non-epistemic modal operators, or in semantic models that represent the interaction of epistemic and non-epistemic concepts. In game theory, for example, an adequate model of the playing of a game must represent not just the epistemic possibilities for each of the players, but also the capacities of players to make each of the choices that are open to that player, even when it is known that the player will not make some of those choices. One might assume that it is common knowledge that Alice will act rationally in a certain game, and it might be that it is known that Alice would be acting irrationally if she chose option X. Nevertheless, it would distort the representation of the game to deny that Alice has the option of choosing action X, and the counterfactual possibility in which she exercises that option may play a role in the deliberations of both Alice and the other players, whose knowledge that Alice will not choose option X is based on their knowledge of what she knows would happen if she did. So even if one makes the idealizing assumption that all agents have only true beliefs, or that it is common belief that everyone's beliefs are true, one should recognize the more general structure that distinguishes belief from knowledge, and that distinguishes both of these concepts from subjective indistinguishability. In the more general structure that recognizes these distinctions, the epistemic accessibility relation is underdetermined by the other relations.

5. MINIMAL AND MAXIMAL EXTENSIONS

So our task is to say more about how to extend the relation D of doxastic accessibility to a relation R of epistemic

accessibility. We know, from the assumption that knowledge implies belief, that in any model meeting our basic conditions on the relation between knowledge and belief, R will be an extension of D (for all x and y, if xDy, then xRy), and we know from the assumption that knowledge implies truth that the extension will be to a reflexive relation. We know by the assumption that belief is strong belief (belief that one knows) that R coincides with D, within the D-set (for all x and y, if xDx, then xRy if and only if xDy). What remains to be said is what determines, for a possible world x that is *outside* of a D-set, which other possible worlds outside that D-set are epistemically accessible to x. If some of my beliefs about what I know are false, what can be said about other propositions that I think that I know?

The assumptions of the neutral theory put clear upper and lower bounds on the answer to this question, and two ways to specify R in terms of the other resources of the model are to make the minimal or maximal extensions. The *minimal* extension of D would be the reflexive closure of D. On this account, the set of epistemically possible worlds for a knower in world x will be the set of doxastically accessible worlds, plus x. To make this minimal extension is to adopt the true belief analysis of knowledge, or in case one is making the internalist assumptions about justified belief, it would be to adopt the justified true belief analysis. The logic of true belief, S4.4, is stronger than S4.2, but weaker than S5.[16] The true belief analysis has its defenders, but most will want to impose stronger conditions on knowledge, which in our setting means that we need to go beyond the minimal extension of R.

It follows from the positive and negative introspection conditions for belief that for any possible world x, all worlds epistemically accessible to x will be subjectively indistinguishable from x (for all x and y, if xRy, then xEy) and this sets the upper bound on the extension of D to R. To identify R with the *maximal* admissible extension is to define it as follows: $xRy =_{df}$ either (xDx and xDy) or (not xDx and xEy). This account of knowledge allows one to know things that go beyond one's internal states only when *all* of one's beliefs are

correct. The logic of this concept of knowledge, S4F, is stronger than S4.2, but weaker than the logic of the minimal extension, S4.4. The maximal extension would not provide a plausible account of knowledge in general, but it might be the appropriate idealization for a certain limited context. Suppose your information all comes from a single source (an oracle), who is presumed, justifiably, to be reliable. Assuming that all of its pronouncements are true, they give you knowledge, but in possible worlds in which any one of its pronouncements is false, it is an unreliable oracle, and so nothing it says should be trusted. This logic, S4F, has been used as the underlying logic of knowledge in some theoretical accounts of a nonmonotonic logic. Those accounts don't provide an intuitive motivation for using this logic, but I think a dynamic model, with changes in knowledge induced by a single oracle who is presumed to be reliable, can provide a framework that makes intuitive sense of these nonmonotonic theories.[17]

6. BELIEF REVISION AND THE DEFEASIBILITY ANALYSIS

Any attempt to give an account of the accessibility relation for knowledge that falls between the minimal and maximal admissible extensions of the accessibility relation for belief will have to enrich the resources of the theory. One way to do this, a way that fits with one of the familiar strategies for responding to the Gettier counterexamples to the justified true belief analysis, is to add to the semantics for belief a theory of belief revision, and then to define knowledge as belief (or justified belief) that is stable under any potential revision by a piece of information that is in fact true. This is the defeasibility strategy followed by many of those who responded to Gettier's challenge: the idea was that the fourth condition (to be added to justified true belief) should be a requirement that there be no "defeater" – no true proposition that, if the knower learned that it was true, would lead her to give up the belief, or to be no longer justified in holding it.[18] There was much discussion in the post-Gettier literature,

about exactly how defeasibility should be defined, but in the context of our idealized semantic models, supplemented by a semantic version of the standard belief revision theory, a formulation of a defeasibility analysis of knowledge is straightforward. First, let me sketch the outlines of the so-called AGM theory of belief revision,[19] and then give the defeasibility analysis.

The belief revision project is to define, for each belief state (the prior belief state), a function taking a proposition (the potential new evidence) to a posterior belief state (the state that would be induced in one in the prior state by receiving that information as one's total new evidence). If belief states are represented by sets of possible worlds (the doxastically accessible worlds), and if propositions are also represented by sets of possible worlds, then the function will map one set of worlds (the prior belief set) to another (the posterior belief set), as a function of a proposition. Let **B** be the set representing the prior belief state, ϕ the potential new information, and $\mathbf{B}(\phi)$ the set representing the posterior state. Let **E** be a superset of **B** that represents the set of all possible worlds that are potential candidates to be compatible with some posterior belief state. The formal constraints on this function are then as follows: (1) $\mathbf{B}(\phi) \subseteq \phi$ (the new information is believed in the posterior belief state induced by that information). (2) If $\phi \cap \mathbf{B}$ is nonempty, then $\mathbf{B}(\phi) = \phi \cap \mathbf{B}$ (If the new information is compatible with the prior beliefs, then nothing is given up – the new information is simply added to the prior beliefs.). (3) $\mathbf{B}(\phi)$ is nonempty if and only if $\phi \cap \mathbf{E}$ is non-empty (the new information induces a consistent belief state whenever that information is compatible with the knower being in the prior belief state. and only then). (4) If $\mathbf{B}(\phi) \cap \psi$ is nonempty, then $\mathbf{B}(\phi \cap \psi) = \mathbf{B}(\phi) \cap \psi$. The fourth conditions is the only one that is not straightforward. What it says is that if ψ is compatible, not with Alice's *prior* beliefs, but with the *posterior* beliefs that she would have if she learned ϕ, then what Alice should believe upon learning the *conjunction* of ϕ and ψ should be the same as what she would believe if she first learned ϕ, and then learned ψ. This condition can be

seen as a generalization of condition (2), which is a modest principle of methodological conservativism (Don't give up any beliefs if your new information is compatible with everything you believe) It is also a kind of path independence principle. The order in which Alice receives two compatible pieces of information should not matter to the ultimate belief state.[20]

To incorporate the standard belief revision theory into our models, add, for each possible world x, and for each agent i, a function that, for each proposition ϕ, takes i's belief state in x, $B_{x,i}=\{y : xD_iy\}$, to a potential posterior belief state, $B_{x,i}(\phi)$. Assume that each of these functions meets the stated conditions, where the set **E**, for the function $B_{x,i}$ is the set of possible worlds that are subjectively indistinguishable from x to agent i. We will also assume that if x and y are subjectively indistinguishable to i, then i's belief revision function will be the same in x as it is in y. This is to extend the positive and negative introspection assumptions to the agent's belief revision policies. Just as she knows what she believes, so she knows how she is disposed to revise her beliefs in response to unexpected information.[21]

We have added some structure to the models, but not yet used it to interpret anything in the object language that our models are interpreting. Suppose our language has just belief operators (and not knowledge operators) for our agents, and only a doxastic accessibility relation, together with the belief revision structure, in the semantics The defeasibility analysis suggests that we might add, for knower i, a knowledge operator with the following semantic rule: $K_i\phi$ is true in world x if $B_i\phi$ is true in x, and for any proposition ψ that is true in x, $\boldsymbol{B}_{x,i}(\psi)\subseteq\phi$. Alice knows that ϕ if and only if, for any ψ that is true, she would still believe that ϕ after learning that ψ. Equivalently, we might define an epistemic accessibility relation in terms of the belief revision structure, and use it to interpret the knowledge operator in the standard way. Let us say that xR_iy if and only if there exists a proposition ϕ such that $\{x,y\}\subseteq\phi$, and $y \in \mathbf{B}_{x,i}(\phi)$. The constraints imposed on the function $\mathbf{B}_{x,i}$ imply that this relation will extend the

doxastic accessibility relation D_i, and that it will fall between our minimal and maximal constraints on this extension. The relation will be transitive, reflexive, and strongly convergent, and so meet all the conditions of our basic theory. It will also meet an additional condition: it will be weakly connected (if xRy and xRz, then either yRz, or zRy). This defeasibility semantics will validate a logic of knowledge, S4.3, that is stronger than S4.2, but weaker than either S4F or S4.4.[22]

So a nice, well behaved version of our standard semantics for knowledge falls out of the defeasibility analysis, yielding a determinate account, in terms of the belief revision structure, of the way that epistemic accessibility extends doxastic accessibility. But I doubt that this is a plausible account of knowledge in general, even in our idealized setting. The analysis is not so demanding as the S4F theory, but like that theory, it threatens to let any false belief defeat too much of our knowledge, even knowledge of facts that seem unrelated. Consider the following kind of example: Alice takes herself to know that the butler didn't do it, since she saw him in the drawing room, miles away from the scene of the crime, at the time of the murder (or so she thinks). She also takes herself to know there is zucchini planted in the garden, since the gardener always plants zucchini, and she saw the characteristic zucchini blossoms on the vines in the garden (or so she thinks). As it happens, the gardener, quite uncharacteristically, failed to plant the zucchini this year, and coincidentally, a rare weed with blossoms that resemble zucchini blossoms has sprung up in its place. But it really was the butler that Alice saw in the drawing room, just as she thought. Does the fact that her justified belief about the zucchini is false take away her knowledge about the butler? It is a fact that either it wasn't really the butler in the drawing room, or the gardener failed to plant zucchini. Were Alice to learn just this disjunctive fact, she would have no basis for deciding which of her two independent knowledge claims was the one that was wrong. So it seems that, on the simple defeasibility account, the disjunctive fact is a defeater. The fact that she is wrong about one of her knowledge claims seems to infect

other, seemingly unrelated claims. Now it may be right that if Alice was in fact reliably informed that one of her two knowledge claims was false, without being given any information about which, she would *then* no longer know that it was the butler that she saw. But if the mere fact that the disjunction is true were enough to rob her of her knowledge about the butler, then it would seem that almost all of Alice's knowledge claims will be threatened. The defeasibility account is closer than one might have thought to the maximally demanding S4F analysis, according to which we know nothing except how things seem to us unless we are right about everything we believe.

I think that one might plausibly defend the claim that the defeasibility analysis provides a *sufficient* condition for knowledge (in our idealized setting), and so the belief revision structure might further constrain the ways in which the doxastic accessibility relation can be extended to an epistemic accessibility relation. But it does not seem to be a plausible *necessary* and sufficient condition for knowledge. In a concluding section, I will speculate about some other features of the relation between a knower and the world that may be relevant to determining which of his true beliefs count as knowledge.

7. THE CAUSAL DIMENSION

What seems to be driving the kind of counterexample to the defeasibility analysis that I have considered is the fact that, on this analysis, a belief with a normal and unproblematic causal source could be defeated by the fact that some different source had delivered misinformation about some independent and irrelevant matter. Conditions were normal with respect to the explanation of Alice's beliefs about the butler's presence in the drawing room. There were no anomalous circumstances, either in her perceptual system, or in the conditions in the environment, to interfere with the normal formation of that belief. This was not the case with respect to the explanation of her belief about what was planted in the garden, but that does

not seem, intuitively, to be relevant to whether her belief about the butler constituted knowledge. Perhaps the explanation of epistemic accessibility, in the case where conditions are not fully normal, and not all of the agent's beliefs are true, should focus more on the causal sources of beliefs, rather than on how agents would respond to information that they do not in fact receive. This, of course, is a strategy that played a central role in many of the responses to the Gettier challenge. I will describe a very simple model of this kind, and then mention some of the problems that arise in making the simple model even slightly more realistic.

Recall that we can formulate the basic theory of belief this way: a relation of subjective indistinguishability, for each agent, partitions the space of possibilities, and there will be a nonempty subset of each partition cell which is the set of worlds compatible with what the agent believes in the worlds in that cell. We labeled those worlds the normal ones, since they are the worlds in which everything determining the agent's beliefs is functioning normally. All of the beliefs are true in those worlds, and belief and knowledge coincide. The problem was to say what the agent knows in the worlds that lie outside of the normal set. One idea is to give a more detailed account of the normal conditions in terms of the way the agent interacts with the world he knows about; we start with a crude and simple model of how this might be done. Suppose our agent receives his information from a fixed set of independent sources – different informants who send messages on which the agent's knowledge is based. The "informants" might be any kind of input channel. The agent might or might not be in a position to identify or distinguish different informants. But we assume that the informants are, in fact, independent in the sense that there may be a fault or corruption that leads one informant to send misinformation (or more generally, to be malfunctioning) while others are functioning normally. So we might index normal conditions to the informant, as well as to the agent. For example, if there are two informants, there will be a set of worlds that is normal with respect to the input channel for informant one,

and an overlapping set that is normal for informant two. Possible worlds in which conditions are fully normal will be those in which all the input channels are functioning normally – the worlds in the intersection of the two sets.[23] This intersection will be the set compatible with the agent's beliefs, the set where belief and knowledge coincide. If conditions are abnormal with respect to informant one (if that information channel is corrupted) then while that informant may influence the agent's beliefs, it won't provide any knowledge. But if the other channel is uncorrupted, the beliefs that have it as their sole source will be knowledge. The formal model suggested by this picture is a simple and straightforward generalization of the S4F model, the maximal admissible extension of the doxastic accessibility relation. Here is a definition of the epistemic accessibility relation for the S4F semantics, where $\mathbf{E}(x)$ is the set of worlds subjectively indistinguishable from x (to the agent in question) and $\mathbf{N}(x)$ is the subset of that set where conditions are normal (the worlds compatible with what the agent believes in world x): xRy if and only if $x \in \mathbf{N}(x)$ and $y \in \mathbf{N}(x)$, or $x \notin \mathbf{N}(x)$ and $y \in \mathbf{E}(x)$. In the generalization, there is a finite set of normal-conditions properties, $\mathbf{N}^j$, one for each informant j, that each determines a subset of $\mathbf{E}(x)$, $\mathbf{N}^j(x)$, where conditions are functioning normally in the relation between that informant and the agent. The definition of R will say that the analogue of the S4F condition holds for each $\mathbf{N}^j$. The resulting logic (assuming that the number of independent information channels or informants is unspecified) will be the same as the basic theory: S4.2.

Everything goes smoothly if we assume that information comes from discrete sources, even if the agent does not identify or distinguish the sources. Even when the agent makes inferences from beliefs derived from multiple sources, some of which may be corrupt and other not, the model will determine which of his true beliefs count as knowledge, and which do not. But in even a slightly more realistic model, the causal explanations for our beliefs will be more complex, with different sources not wholly independent, and deviations from normal conditions hard to isolate. Beliefs may have multiple

interacting sources – there will be cases of overdetermination and preemption; there will be problems about how to treat cases where a defect in the system results, not in the reception of misinformation, but in the failure to receive a message. (It might be that had the system been functioning normally, I would have received information that would have led me to give up a true belief.) And along with complicating the causal story, one might combine this kind of model with a belief revision structure, allowing one to explore the relation between beliefs about causal structure and policies for belief revision, and to clarify the relation between the defeasibility analysis and an account based on the causal strategy. The abstract problems that arise when one tries to capture a more complex structure will reflect, and perhaps help to clarify, some of the patterns in the counterexamples that arose in the post-Gettier literature. Our simple model abstracts away from most of these problems, but it is a start that may help to provide a context for addressing them.

APPENDIX

To give a very concise summary of all the logics of knowledge I have discussed, and their corresponding semantics, I will list, first the alternative constraints on the accessibility relation, and then the alternative axioms. Then I will distinguish the different logics, and the semantic conditions that are appropriate to them in terms of the items on the lists.

Conditions on R:	
(Ref)	$(x)xRx$
(Tr)	$(x)(y)(z)((xRy \ \& \ yRz) \rightarrow xRz)$
(Cv)	$(x)(y)(z)((xRy \ \& \ xRz) \rightarrow (\exists w)(yRw \ \& \ zRw))$
(SCv)	$(x)(\exists z)(y)(xRy \rightarrow yRz)$
(WCt)	$(x)(y)(z)((xRy \ xRz) \rightarrow (yRz \ \ zRy))$
(F)	$(x)(y)(xRy \rightarrow ((z)(xRz \rightarrow yRz) \ \ (z) \rightarrow (xRz \rightarrow zRy))$
(TB)	$(x)(y)((xRy \ \& \ x \neq y) \rightarrow (z)(xRz \rightarrow zRy))$
(E)	$(x)(y)(z)((xRy \ \& \ xRz) \rightarrow yRz)$

Axioms:	
(T)	$K\phi \rightarrow \phi$
(4)	$K\phi \rightarrow \mathrm{KK}\phi$
(4.2)	$\mathrm{MK}\phi \rightarrow \mathrm{KM}\phi$
(4.3)	$(K(\phi \rightarrow M\psi) \quad K(\psi \rightarrow M\phi))$
(f)	$((M\phi \ \& \ \mathrm{MK}\psi) \rightarrow K(M\phi \vee \psi))$
(4.4)	$((\phi \ \& \ \mathrm{MK}\psi) \rightarrow K(\phi \vee \psi))$
(5)	$M\phi \rightarrow \mathrm{KM}\phi$

The logics for knowledge we have considered, and semantic constraints on R relative to which they are sound and complete, are as follows. The logics are of increasing order of strength, the theorems of each including those of the previous logics on the list.

S4	K + T + 4	Ref + Tr		
S4.2	S4 + 4.2	Ref + Tr + SCv	OR	Ref + Tr + Cv
S4.3	S4 + 4.3	Ref + Tr + SCv + WCt	OR	Ref + Tr + WCt
S4F	S4 + f	Ref + Tr + F		
S4.4	S4 + 4.4	Ref + Tr + TB		
S5	S4 + 5	Ref + Tr + E		

In each of the logics of knowledge we have considered, from S4.2 to S4.4, the derived logic of belief, with belief defined by the complex operator MK, will be KD45. (In S4, belief is not definable, since in that logic, the complex operator MK does not satisfy the K axiom, and so is not a normal modal operator. In S5, belief and knowledge coincide, so the logic of belief is S5.) KD45 is $K + D + 4 + 5$, where D is $(K\phi \rightarrow M\phi)$. The semantic constraints are Tr + E + the requirement that the accessibility relation be serial: $(x)(\exists y)xRy$.

In a semantic model with multiple knowers, we can add a common knowledge operator, defined in terms of the transitive closure of the epistemic accessibility relations for the different knowers. For any of the logics, from S4 to S4.4, with the corresponding semantic conditions, the logic of common knowledge will be S4, and the accessibility relation will be

transitive and reflexive, but will not necessarily have any of the stronger properties. If the logic of knowledge is S5, then the logic of common knowledge will also be S5, and the accessibility relation will be an equivalence relation.

NOTES

[1] See Fagin et al. (1995) and Battigalli and Bonanno (1999) for excellent surveys of the application of logics of knowledge and belief in theoretical computer science and game theory.

[2] I explore the problem of logical omniscience in two papers, Stalnaker (1991) and (1999b). I don't attempt to solve the problem in either paper, but only to clarify it, and to argue that it is a genuine problem, and not an artifact of a particular theoretical framework.

[3] Substituting '$\sim K\phi$ ' for ψ, and eliminating a double negation, the principle says that if $\{K\phi, \sim KK\phi\}$ is consistent, then $\{K\phi, \sim K\phi\}$ is consistent.

[4] See especially, Williamson (2000) for some reasons to reject the KK principle.

[5] Hintikka (1962), 106.

[6] *Ibid*, 54.

[7] More precisely , if R^i is the accessibility relation for knower i, then the common-knowledge accessibility relation for a group G is defined as follows; xR^Gy if there is a sequence of worlds, $z_1,...z_n$ such that $z_1 = x$ and $z_n = y$ and for all j between 1 and $n-1$, there is a knower $i \in G$, such that $z_jR^i\ z_{j+1}$.

[8] A more complex kind of model would specify a set of admissible *initial* global states, and a set of transition rules taking global states to global states. The possible worlds in this kind of model are the admissible global *histories* – the possible ways that the system might evolve. In this kind of model, one can represent the distribution of information, not only about the current state of the system, but also about how it evolved, and where it is going. In the more general model, knowledge states are time-dependent, and the components may have or lack information not only about which possible world is actual, but also about where (temporally) it is in a given world. The dynamic dimension, and the parallels with issues about indexical knowledge and belief, are part of the interest of the distributed systems models, but I will ignore these issues here.

[9] Possible worlds, on this way of formulating the theory, are not primitive points, as they are in the usual abstract semantics, but complex objects – sequences of local states. But an equivalent formulation might begin with a given set of primitive (global) states, together with a set of

equivalence relations, one for each knower, and one for "nature". The local states could then be defined as the equivalence classes.

[10] These issues are discussed in Stalnaker (1996).

[11] KD45 adds to the basic modal system *K* the axioms (*D*), which is our CB, (4) $B\phi \rightarrow \mathrm{BB}\phi$, which follows immediately from our (PI) and (KB), and (5) $\sim B\phi \rightarrow B\sim B\phi$, which follows immediately from (NI) and (KB). The necessitation rule for *B* (If $\vdash\phi$, then $\mathrm{B}\phi$) and the distribution principle ($B(\phi \rightarrow \psi) \rightarrow (B\phi \rightarrow B\psi)$) can both be derived from our principles.

[12] The definability of belief in terms of knowledge, and the point that the assumptions about the relation between knowledge and belief imply that the logic of knowledge should be S4.2, rather than S4, were first shown by Wolfgang Lenzen. See his classic monograph, *Recent Work in Epistemic Logic. Acta Philosophica Fennica* **30** (1978). North-Holland, Amsterdam.

[13] The difference between strong and weak convergence does not affect the *propositional* modal logic, but it will make a difference to the quantified modal logic. The following is an example of a sentence that is valid in models satisfying strong convergence (along with transitivity and reflexivity) but not valid in all models satisfying weak convergence: $\mathrm{MK}((x)(\mathrm{MK}\phi \rightarrow \phi))$

[14] We observed in Note 7 that an equivalent formulation of the S5 distributed systems models would take the global world-states as primitive, specifying an equivalence relation for each agent, and defining local states as equivalence classes of global states. In an equivalent formulation of this kind of the more general theory, the assumption that every sequence of local states is a possible world will be expressed by a recombination condition: that for every sequence of equivalence classes (one for each agent) there is a possible world that is a member of their intersection. I have suggested that a recombination condition of this kind should be imposed on game theoretic models (where the equivalence classes are types, represented by probability functions), defending it as a representation of the conceptual independence of the belief states of different agents.

[15] In most formulations of a possible-worlds semantics for propositional modal logic, a *frame* consists simply of a set of worlds and an accessibility relation. A model on a frame determines the truth values of sentences, relative to each possible world. On this conception of a model, one cannot talk of the truth of a sentence in a model, but only of truth *at a world* in a model. Sentence validity is defined, in formulations of this kind, as truth in all worlds in all models. But in some formulations, including in Kripke's original formal work, a frame (or model structure, as Kripke called it at the time) included, in addition to a set of possible worlds and an accessibility relation, a designated possible world – the actual world of the model. A sentence is true in a model if it is true in the designated actual world, and

valid if true in all models. This difference in formulation was a minor detail in semantic theories for most of the normal modal logics, since any possible world of a model might be the designated actual world without changing anything else. So the two ways of defining sentence validity will coincide. But the finer-grained definition of a frame allows for theories in which the constraints on R, and the semantic rules for operators, make reference to the actual world of the model. In such theories, truth in all worlds in all models may diverge from truth in all models, allowing for semantic models of logics that fail to validate the rule of necessitation.

[16] See the appendix for a summary of all the logics of knowledge discussed, their semantics, and the relationships between them.

[17] See Schwarz and Truszczyski (1992).

[18] See Lehrer and Paxson (1969) and Swain (1974) for two examples.

[19] See Gärdenfors (1988) for a survey of the basic ideas of the AGM belief revision theory, and Grove (1988) for a semantic formulation of the theory.

[20] The third principle is the least secure of the principles; there are counterexamples that suggest that it should be given up. See Stalnaker (1994) for a discussion of one. The defeasibility analysis of knowledge can be given with either the full AGM belief revision theory, or with the more neutral one that gives up the fourth condition.

[21] It should be noted that even with the addition of the belief revision structure to the epistemic models I have been discussing, they remain static models. A model of this kind represents only the agents' beliefs at a fixed time, together with the policies or dispositions to revise her beliefs that she has at that time. The model does not represent any actual revisions that are made when new information is actually received. The models can be enriched by adding a temporal dimension to represent the dynamics, but doing so requires that the knowledge and belief operators be time indexed, and that one be careful not to confuse belief changes that are changes of mind with belief changes that result from a change in the facts. (I may stop believing that the cat is on the mat because I learn that what I thought was the cat was the dog, or I may stop believing it because the cat gets up and leaves, and the differences between the two kinds of belief change are important)

[22] In game theoretic models, the strength of the assumption that there is common knowledge of rationality depends on what account one gives of knowledge (as well as on how one explains rationality). Some backward induction arguments, purporting to show that common knowledge of rationality suffices to determine a particular course of play (in the centipede game, or the iterated prisoners' dilemma, for example) can be shown to work with a defeasibility account of knowledge, even if they fail on a more neutral account. See Stalnaker (1996).

[23] It will be required that the intersection of all the normal-conditions sets be nonempty.

REFERENCES

Battigalli, P. and Bonanno, G. (1999): 'Recent Results on Belief, Knowledge and the Epistemic Foundations of Game Theory', *Research in Economics* 53, 149–225.

Fagin, R., Halpern, J., Moses, Y. and Vardi, M. (1995): *Reasoning about Knowledge*, Cambridge, MA: MIT Press.

Gärdenfors, P. (1988): *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, Cambridge, MA: MIT Press.

Gettier, E. (1963): 'Is Justified True Belief Knowledge?', *Analysis* 6, 121–123.

Grove, A. (1988): 'Two Modeling for Theory Change', *Journal of Philosophical Logic* 17, 157–170.

Hintikka, J. (1962): *Knowledge and Belief*, Ithaca, NY: Cornell University Press.

Lehrer, K. and Paxson, T. (1969): 'Knowledge: Undefeated Justified True Belief', *The Journal of Philosophy* 66, 225–237.

Lenzen, W. (1978): *Recent Work in Epistemic Logic*, 30, Amsterdam, North-Holland: Acta Philosophica Fennica.

Schwarz, G. and Truszczyński, M. (1992): 'Modal Logic S4F and the Minimal Knowledge Paradigm', *Proceedings of the Fourth Conference on Theoretical Aspects of Reasoning about Knowledge* (pp. 184–198), Sam Mateo, CA: Morgan Kaufmann Publishers, Inc.

Stalnaker, R. (1991): 'The Problem of Logical Omniscience, I', *Synthese* 89, 425–440 (Reprinted in Stalnaker, (1999a), 240–254.

Stalnaker, R. (1994): 'What is a Non-monotonic Consequence Relation?', *Fundamenta Informaticae* 21, 7–21.

Stalnaker, R. (1996): 'Knowledge, Belief and Counterfactual Reasoning in Games', *Economics and Philosophy* 12, 133–162.

Stalnaker, R. (1999a): *Context and Content: Essays on Intentionality in Speech and Thought*, Oxford: Oxford University Press.

Stalnaker, R. (1999b): 'The Problem of Logical Omniscience II', in Stalnaker, (1999a), 255–273.

Swain, M. (1974): 'Epistemic Defeasibility', *The American Philosophical Quarterly* 11, 15–25.

Williamson, T. (2000): *Knowledge and Its Limits*, Oxford: Oxford University Press.

Department of Linguistics and Philosophy
Massachusetts Institute of Technology
77 Massachussetts Avenue, 32D-808
Cambridge, MA 02139
USA
E-mail: stal@mit.edu