Chapter 42
Knowledge, Belief and Counterfactual Reasoning in Games

Robert Stalnaker

Introduction

Deliberation about what to do in any context requires reasoning about what will or would happen in various alternative situations, including situations that the agent knows will never in fact be realized. In contexts that involve two or more agents who have to take account of each others' deliberation, the counterfactual reasoning may become quite complex. When I deliberate, I have to consider not only what the causal effects would be of alternative choices that I might make, but also what other agents might believe about the potential effects of my choices, and how their alternative possible actions might affect my beliefs. Counterfactual possibilities are implicit in the models that game theorists and decision theorists have developed – in the alternative branches in the trees that model extensive form games and the different cells of the matrices of strategic form representations – but much of the reasoning about those possibilities remains in the informal commentary on and motivation for the models developed. Puzzlement is sometimes expressed by game theorists about the relevance of what happens in a game 'off the equilibrium path': of what would happen if what is (according to the theory) both true and known by the players to be true were instead false. My aim in this paper is to make some suggestions for clarifying some of the concepts involved in counterfactual reasoning in strategic contexts, both the reasoning of the rational agents being modeled, and the reasoning of the theorist who is doing the modeling, and to bring together some ideas and technical tools developed by philosophers and logicians that I think might be relevant to the analysis of strategic reasoning, and more generally to the conceptual foundations of game theory.

R. Stalnaker (✉)
Department of Linguistics and Philosophy, MIT, Cambridge, MA, USA
e-mail: stal@mit.edu

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate
Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_42

There are two different kinds of counterfactual possibilities – causal and epistemic possibilities – that need to be distinguished. They play different but interacting roles in a rational agent's reasoning about what he and others will and should do, and I think equivocation between them is responsible for some of the puzzlement about counterfactual reasoning. In deliberation, I reason both about how the world might have been different if I or others did different things than we are going to do, and also about how my beliefs, or others' beliefs, might change if I or they learned things that we expect not to learn. To take an often cited example from the philosophical literature to illustrate the contrast between these two kinds of counterfactual suppositions, compare: *if Shakespeare didn't write Hamlet, someone else did,* with *if Shakespeare hadn't written Hamlet, someone else would have*.[1] The first expresses a quite reasonable disposition to hold onto the belief that someone wrote Hamlet should one receive the unexpected information that Shakespeare did not; the second expresses a causal belief, a belief about objective dependencies, that would be reasonable only if one held a bizarre theory according to which authors are incidental instruments in the production of works that are destined to be written. The content of what is supposed in the antecedents of these contrasting conditionals is the same, and both suppositions are or may be counterfactual in the sense that the person entertaining them believes with probability one that what is being supposed is false. But it is clear that the way it is being supposed is quite different in the two cases.

This contrast is obviously relevant to strategic reasoning. Beliefs about what it is rational to do depend on causal beliefs, including beliefs about what the causal consequences would be of actions that are alternatives to the one I am going to choose. But what is rational depends on what is believed, and I also reason about the way my beliefs and those of others would change if we received unexpected information. The two kinds of reasoning interact, since one of the causal effects of a possible action open to me might be to give unexpected information to another rational agent.[2]

It is obvious that a possible course of events may be causally impossible even if it is epistemically open, as when you have already committed yourself, but I have not yet learned of your decision. It also may happen that a course of events is causally open even when it is epistemically closed in the sense that someone believes, with probability one, that it will not happen. But can it be true of a causally open course of events that someone not only believes, but also *knows* that it will not occur? This is less clear; it depends on how we understand the concept of knowledge. It does not seem incoherent to suppose that you know that I am rational, even though irrational choices are still causally possible for me. In fact, the concept of rationality seems applicable to actions only when there are options open to an agent. If we are to make sense of assumptions of knowledge and common knowledge of rationality, we need

[1] Ernest Adams (1970) first pointed to the contrast illustrated by this pair of conditionals. The particular example is Jonathan Bennett's.

[2] The relation between causal and evidential reasoning is the central concern in the development of causal decision theory. See Gibbard and Harper (1981), Skyrms (1982) and Lewis (1980).

to allow for the possibility that an agent may know what he or another agent is going to do, even when it remains true that the agent could have done otherwise.

To clarify the causal and epistemic concepts that interact in strategic reasoning, it is useful to break them down into their component parts. If, for example, there is a problem about exactly what it means to assume that there is common knowledge of rationality, it ought to be analyzed into problems about exactly what rationality is, or about what knowledge is, or about how common knowledge is defined in terms of knowledge. The framework I will use to represent these concepts is one that is designed to help reveal the compositional structure of such complex concepts: it is a formal semantic or model theoretic framework – specifically, the Kripkean 'possible worlds' framework for theorizing about modal, causal and epistemic concepts. I will start by sketching a simple conception of a model, in the model theorist's sense, of a strategic form game. Second, I will add to the simple conception of a model the resources to account for one kind of counterfactual reasoning, reasoning about belief revision. In these models we can represent concepts of rationality, belief and common belief, and so can define the complex concept of common belief in rationality, and some related complex concepts, in terms of their component parts. The next step is to consider the concept of knowledge, and the relation between knowledge and belief. I will look at some different assumptions about knowledge, and at the consequences of these different assumptions for the concepts of common knowledge and common knowledge of rationality. Then to illustrate the way some of the notions I discuss might be applied to clarify some counterfactual reasoning about games, I will discuss some familiar problems about backward induction arguments, using the model theory to sharpen the assumptions of those arguments, and to state and prove some theorems about the consequences of assumptions about common belief and knowledge.

Model Theory for Games

Before sketching the conception of a model of a game that I will be using, I will set out some assumptions that motivate it, assumptions that i think will be shared by most though not all, game theorists. First, I assume that a game is a *partial* description of a set or sequence of interdependent Bayesian decision problems. The description is partial in that while it specifies all the relevant utilities motivating the agents, it does not give their degrees of belief. Instead, qualitative constraints are put on what the agents are assumed to believe about the actions of other agents; but these constraints will not normally be enough to determine what the agents believe about each other, or to determine what solutions are prescribed to the decision problems. Second, I assume that all of the decision problems in the game are problems of individual decision making. There is no special concept of rationality for decision making in a situation where the outcomes depend on the actions of more than one agent. The acts of other agents are, like chance events, natural disasters and acts of God, just facts about an uncertain world that agents have beliefs and degrees of belief about. The utilities of other agents are relevant to an agent only as information that, together with beliefs about the rationality of those agents, helps to predict their actions. Third, I assume that in cases where degrees of belief are undetermined, or only partially determined, by the description of a decision problem, then no action is prescribed by the theory unless there is an action that would be rational for every system of degrees of belief compatible with what is specified. There are no special rules of rationality telling one what to do in the absence of degrees of belief, except this: decide what you believe, and then maximize expected utility.

A model for a game is intended to represent a completion of the partial specification of the set or sequence of Bayesian decision problems that is given by the definition of the game, as well as a representation of a particular play of the game. The class of all models for a game will include all ways of filling in the relevant details that are compatible with the conditions imposed by the definition of the game. Although a model is intended to represent one particular playing of the game, a single model will contain many possible worlds, since we need a representation, not only of what actually happens in the situation being modeled, but also what might or would happen in alternative situations that are compatible with the capacities and beliefs of one or another of the agents. Along with a set of possible worlds, models will contain various relations and measures on the set that are intended to determine all the facts about the possible worlds that may be relevant to the actions of any of the agents playing the game in a particular concrete context.

The models considered in this paper are models for finite games in normal or strategic form. I assume, as usual, that the game Γ itself consists of a structure $\langle N, \langle C_i, u_i \rangle_{i \in N} \rangle$, where N is a finite set of players, C_i is a finite set of alternative strategies for player i, and u_i is player i′s utility function taking a strategy profile (a specification of a strategy for each player) into a utility value for the outcome that would result from that sequence of strategies. A model for a game will consist of a set of possible worlds (a state space), one of which is designated as the actual world of the model. In each possible world in the model, each player has certain beliefs and partial beliefs, and each player makes a certain strategy choice. The possible worlds themselves are simple, primitive elements of the model; the information about them – what the players believe and do in each possible world – is represented by several functions and relations given by a specification of the particular model. Specifically, a model for a game Γ will consist of a structure $\langle W, a, \langle S_i, R_i, P_i \rangle_{i \in N} \rangle$, where W is a nonempty set (the possible worlds), a is a member of W (the actual world), each S_i is a function taking possible worlds into strategy choices for player i, each R_i is a binary relation on W, and each P_i is an additive measure function on subsets of W.

The R relations represent the qualitative structure of the players' beliefs in the different possible worlds in the following way: the set of possible worlds that are compatible with what player i believes in world w is the set $\{x : wR_ix\}$. It is assumed that the R relations are *serial, transitive,* and *euclidean*.[3] The first assumption is simply the requirement that in any possible world there must be at least one possible world compatible with what any player believes in that world. The other two constraints encode the assumption that players know their own minds: they are necessary and sufficient to ensure that players have introspective access to their beliefs: if they believe something, they believe that they believe it, and if they do not, they believe that they do not.

The S functions encode the facts about what the players do – what strategies they choose – in each possible world. It is assumed that if xR_iy, then $S_i(x) = S_i(y)$. Intuitively, this requirement is the assumption that players know, at the moment of choice, what they are doing – what choice they are making. Like the constraints on the structure of the R relations, this constraint is motivated by the assumption that players have introspective access to their own states of mind.

The measure function P_i, encodes the information about the player's *partial* beliefs in each possible world in the following way: player i′s belief function in possible world w is the relativization of P_i to the set $\{x{:}wR_ix\}$. That is, for any proposition ϕ, $P_{i,W}(\phi) = P_i(\phi \cap\{x{:}wR_ix\})/P_i(\{x{:}wR_ix\})$. The assumptions we are making about R_i and P_i will ensure that $P_i(\{x{:}wR_ix\})$ is nonzero for all w, so that this probability will always be defined.

The use of a single measure function for each player, defined on the whole space of possible worlds, to encode the information required to define the player's degrees of belief is just a technical convenience – an economical way to specify the many different belief functions that represent that player's beliefs in different possible worlds. No additional assumptions about the players' beliefs are implicit in this form of representation, since our introspection assumptions already imply that any two different belief states for a single player are disjoint, and any set of probability measures on disjoint sets can be represented by a single measure on the union of all the sets. This single measure will contain some extraneous information that has no representational significance – different total measures will determine the same set of belief functions – but this artifact of the model is harmless.[4]

[3] That is, for all players i $(x)(\exists y)xR_iy$, $(x)(y)(z)((xR_iy \ \&\ yR_iz) \rightarrow xR_iz)$, and $(x)(y)(z)((xR_iy \ \&\ xR_iz) \rightarrow yRiz)$.

[4] It has been suggested that there is a substantive, and implausible, assumption built into the way that degrees of belief are modeled: namely, that any two worlds in which a player has the same *full* beliefs he also has the same *partial* beliefs. But this assumption is a tautological consequence of the introspection assumption, which implies that a player fully believes that he himself has the partial beliefs that he in fact has. It does follow from the introspection assumptions that player j cannot be uncertain about player i's partial beliefs while being certain about all of i's full beliefs. But that is just because the totality of i's full beliefs includes his beliefs about his own partial beliefs, and by the introspection assumption, i's beliefs about his own partial beliefs are complete and correct. Nothing, however, prevents there being a model in which there are different worlds in which player i has full beliefs about objective facts that are exactly the same, even though the degrees of belief about such facts are different. This situation will be modeled by disjoint but isomorphic sets of possible worlds. In such a case, another player j might be certain about player i's full beliefs about everything except i's own partial beliefs, while being uncertain about i's partial beliefs.

In order to avoid complications that are not relevant to the conceptual issues I am interested in, I will be assuming throughout this discussion that our models are finite, and that the measure functions all assign nonzero probability to every nonempty subset of possible worlds.

We need to impose one additional constraint on our models, a constraint that is motivated by our concern with counterfactual reasoning. A specification of a game puts constraints on the causal consequences of the actions that may be chosen in the playing of the game, and we want these constraints to be represented in the models. Specifically, in a strategic form game, the assumption is that the strategies are chosen independently, which means that the choices made by one player cannot influence the beliefs or the actions of the other players. One could express the assumption by saying that certain counterfactual statements must be true in the possible worlds in the model: if a player had chosen a different strategy from the one he in fact chose, the other players would still have chosen the same strategies, and would have had the same beliefs, that they in fact had. The constraint we need to add is a closure condition on the set of possible worlds – a requirement that there be enough possible worlds of the right kind to represent these counterfactual possibilities.

For any world w and strategy s for player i, there is a world f(w,s) meeting the following four conditions:

1. for all $j \neq i$, if wR_jx, then $f(w,s)R_jx$.
2. if wR_ix, then $f(w,s)R_if(x,s)$.
3. $S_i(f(w,s)) = s$
4. $P_i(f(w,s)) = P_i(w)$.

Intuitively, f(w,s) represents the counterfactual possible world that, in w, is the world that would have been realized if player i, believing exactly what he believes in w about the other players, had chosen strategy s.

Any of the (finite) models constructed for the arguments given in this paper can be extended to (finite) models satisfying this closure condition. One simply adds, for each $w \in W$ and each strategy profile c, a world corresponding to the pair (w,c), and extending the R's, P's, and S's in a way that conforms to the four conditions.[5]

Because of our concern to represent counterfactual reasoning, it is essential that we allow for the possibility that players have false beliefs in some possible worlds, which means that a world in which they have certain beliefs need not itself be compatible with those beliefs. Because the epistemic structures we have defined allow for false belief, they are more general than the partition structures that will be more familiar to game theorists. An equivalence relation meets the three conditions we have imposed on our R relations, but in addition must be reflexive. To impose this

[5] More precisely, for any given model $M = \langle W,a,\langle S_i,R_i,P_i\rangle_{i\in N}\rangle$, not necessarily meeting the closure condition, define a new model M′ as follows: $W' = W \times C; a' = \langle a,S(a)\rangle$; for all $w \in W$ and $c \in C$, $S'(\langle w,c\rangle) = c$; for all $x,y \in W$ and $c,d \in C$, $\langle x,c\rangle R'_i \langle y,d\rangle$ if the following three conditions are met: (i) xR_iy, (ii) $c_i = d_i$, and (iii) for all $j \neq i$, $S_j(y) = d_j$; $P'_i(\langle x,c\rangle) = P_i(x)$. This model will be finite if the original one was, and will satisfy the closure condition.

additional condition would be to assume that all players necessarily have only true beliefs. But even if an agent in fact has only true beliefs, counterfactual reasoning requires an agent to consider possible situations in which some beliefs are false. First, we want to consider belief contravening, or epistemic counterfactuals: how players would revise their beliefs were they to learn they were mistaken. Second, we want to consider deliberation which involves causal counterfactuals: a player considers what the consequences would be of his doing something he is not in fact going to do. In both cases, a player must consider possible situations in which either she or another player has a false belief.

Even though the R relations are not, in general, equivalence relations, there is a relation definable in terms of R that does determine a partition structure: say that two worlds x and y are *subjectively indistinguishable* for player i ($x \approx_i y$) if player i′s belief state in x is the same as it is in y. That is, $x \approx_i y$ if and only if $\{z{:}xR_iz\} = \{z{:}yR_iz\}$. Each equivalence class determined by a subjective indistinguishability relation will be divided into two parts: the worlds compatible with what the player believes, and the worlds that are not. In the regular partition models, all worlds are compatible with what the player believes in the world, and the two relations, R_i and $\approx_i$, will coincide.

To represent counterfactual reasoning, we must also allow for possible worlds in which players act irrationally. Even if I am resolved to act rationally, I may consider in deliberation what the consequences would be of acting in ways that are not. And even if I am certain that you will act rationally, I may consider how I would revise my beliefs if I learned that I was wrong about this. Even models satisfying some strong condition, such as common belief or knowledge that everyone is rational, will still be models that contain counterfactual possible worlds in which players have false beliefs, and worlds in which they fail to maximize expected utility.

The aim of this model theory is generality: to make, in the definition of a model, as few substantive assumptions as possible about the epistemic states and behavior of players of a game in order that substantive assumptions can be made explicit as conditions that distinguish some models from others. But of course the definition inevitably includes a range of idealizing and simplifying assumptions, made for a variety of reasons. Let me just mention a few of the assumptions that have been built into the conception of a model, and the reasons for doing so.

First, while we allow for irrational action and false belief, we do assume (as is usual) that players all have coherent beliefs that can be represented by a probability function on some nonempty space of possibilities. So in effect, we make the outrageously unrealistic assumption that players are logically omniscient. This assumption is made only because it is still unclear, either conceptually or technically, how to understand or represent the epistemic situations of agents that are not ideal in this sense. This is a serious problem, but not one I will try to address here.

Second, as I have said, it is assumed that players have introspective access to their beliefs. This assumption could be relaxed by imposing weaker conditions on the R relations, although doing so would raise both technical and conceptual problems. It is not clear how one acts on one's beliefs if one does not have introspective access to them. Some may object to the introspective assumption on the ground that a person may have unconscious or inarticulate beliefs, but the assumption is not incompatible with this: if beliefs can be unconscious, so can beliefs about beliefs. It is not assumed that one knows how to say what one believes.

Third, some have questioned the assumption that players know what they do. This assumption might be relaxed with little effect; what is its motivation? The idea is simply that in a static model for a strategic form game, we are modeling the situation at the moment of choice, and it seems reasonable to assume that at that moment, the agent knows what choice is being made.

Fourth, it is assumed that players know the structure of the game – the options available and the utility values of outcomes for all of the players. This assumption is just a simplifying assumption made to avoid trying to do too much at once. It could easily be relaxed with minimal effect on the structure of the models, and without raising conceptual problems. That is, one could consider models in which different games were being played in different possible worlds, and in which players might be uncertain or mistaken about what the game was.

Finally, as noted we assume that models are finite. This is again just a simplifying assumption. Relaxing it would require some small modifications and add some mathematical complications, but would not change the basic story.

In any possible worlds model, one can identify *propositions* with subsets of the set of possible worlds, with what economists and statisticians call 'events'. The idea is to identify the content of what someone may think or say with, its truth conditions – that is, with the set of possible worlds that would realize the conditions that make what is said or thought true. For any proposition ϕ and player i, we can define the proposition that i fully believes that ϕ as the set $\{x \in W: \{y \in W: xR_i y\} \subseteq \phi\}$, and the proposition that i believes that ϕ to at least degree r as the set $\{x \in W: P_{i,x}(\phi) \geq r\}$. So we have the resources to interpret unlimited iterations of belief in any proposition, and the infinitely iterated concept of *common belief* (all players believe that ϕ, and all believe that all believe that ϕ, and all believe that all believe that all believe that ϕ, and ... etc.) can be defined as the intersection of all the propositions in this infinite conjunction. Equivalently, we can represent common belief in terms of the *transitive closure* R*, of the set all the R relations. For any proposition ϕ, it is, in possible world x, common belief among the players that ϕ if and only if ϕ is true in all possible worlds compatible with common belief, which is to say if and only if $\{y: xR^* y\} \subseteq \phi$.

If rationality is identified with maximizing expected utility, then we can define, in any model, the propositions that some particular player is rational, that all players are rational, that all players believe that all players are rational, and of course that it is common belief among the players that all players are rational. Here is a sequence of definitions, leading to a specification of the proposition that there is common belief that all players are rational[6]: first, the *expected utility* of an action (a strategy choice) s for a player i in a world x is defined in the familiar way:

[6] In these and other definitions, a variable for a strategy or profile, enclosed in brackets, denotes the proposition that the strategy or profile is realized. So, for example, if $e \in C_{-i}$ (if e is a strategy profile for players other than player i) then $[e] = \{x \in W: S_j(x) = e_j \text{ for all } j \neq i\}$.

$$\mathrm{eu}_{\mathrm{i,x}}(\mathrm{s}) = \sum_{\mathrm{e}\in\mathrm{C}_{-\mathrm{i}}} P_{i,x}([e]) \times u_i((s,\ e))$$

Second, we define the set of strategies that maximize expected utility for player i in world x:

$$\mathrm{r}_{\mathrm{i,x}} = \left\{\mathrm{s} \in \mathrm{C}_{\mathrm{i}}\ :\ \mathrm{eu}_{\mathrm{i,x}}(\mathrm{s}) \geq \mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}'\right)\ \text{for all } \mathrm{s}' \in\ \mathrm{C}_{\mathrm{i}}\right\}$$

Third, the proposition *that player i is rational* is the set of possible worlds in which the strategy chosen maximizes expected utility in that world:

$$\mathrm{A}_{\mathrm{i}} = \{\mathrm{x} \in \mathrm{W}\ :\ \mathrm{S}_{\mathrm{i}}(\mathrm{x}) \in \mathrm{r}_{\mathrm{i,x}}\}$$

Fourth, the proposition *everyone is rational* is the intersection of the A_{i}'s:

$$\mathrm{A} = \cap_{\mathrm{i}\in\mathrm{N}}\mathrm{A}_{\mathrm{i}}$$

Fifth, the proposition *there is common belief that everyone is rational* is defined as follows:

$$\mathrm{Z} = \{\mathrm{x} \in \mathrm{W}\ :\ \{\mathrm{y} \in \mathrm{W} : \mathrm{xR}^*\mathrm{y}\} \subseteq \mathrm{A}\}.$$

Any specification that determines a proposition relative to a model can also be used to pick out a class of models – all the models in which the proposition is true in that model's actual world. So for any given game, we can pick out the class of models of that game that satisfy some intuitive condition, for example, the class of models in which the proposition Z, that there is common belief in rationality, is true (in the actual world of the model). A class of models defined this way in turn determines a set of strategy profiles for the game: a profile is a member of the set if and only if it is realized in the actual world of one of the models in the class of models. This fact gives us a way that is both precise and intuitively motivated of defining a solution concept for games, or of giving a proof of adequacy for a solution concept already defined. The solution concept that has the most transparent semantic motivation of this kind is rationalizability: we can define rationalizability semantically as the set of strategies of a game that are realized in (the actual world of) some model in which there is common belief in rationality.[7] Or, we can give

[7] This model theoretic definition of rationalizability coincides with the standard concept defined by Bernheim (1984) and Pearce (1984) only in two person games. In the general case, it coincides with the weaker concept, correlated rationalizability. Model theoretic conditions appropriate for the stronger definition would require that players' beliefs about each other satisfy a constraint that (in games with more than two players) goes beyond coherence: specifically, it is required that no player can believe that any information about another player's strategy choices would be evidentially relevant to the choices of a different player. I think this constraint could be motivated, in general, only if one confused causal with evidential reasoning. The structure of the game ensures

a direct nonsemantic definition of the set of strategies – the set of strategies that survive the iterated elimination of strictly dominated strategies – and then prove that this set is *characterized* by the class of models in which there is common belief in rationality: a set of strategies is characterized by a class of models if the set includes exactly the strategies that are realized in some model in the class.[8]

Belief Revision

There are many ways to modify and extend this simple conception of a model of a game. I will consider here just one embellishment, one that is relevant to our concern with counterfactual reasoning. This is the addition of some structure to model the players' policies for revising their beliefs in response to new information. We assume, as is usual, that rational players are disposed to revise their beliefs by conditionalization, but there is nothing in the models we have defined to say how players would revise their beliefs if they learned something that had a prior probability of 0 – something incompatible with the initial state of belief. A belief revision policy is a way of determining the sets of possible worlds that define the posterior belief states that would be induced by such information. The problem is not to generate such belief revision policies out of the models we already have – that is impossible. Rather, it is to say what new structure needs to be added to the model in order to represent belief revision policies, and what formal constraints the policies must obey.

Since we are modeling strategic form games, our models are static, and so there is no representation of any actual change in what is believed. But even in a static situation, one might ask how an agent's beliefs are disposed to change were he to learn that he was mistaken about something he believed with probability one, and the answer to this question may be relevant to his decisions. These dispositions to change beliefs, in contrast to the potential changes that would display the dispositions, are a part of the agent's prior subjective state – the only state represented in the worlds of our models.

I said at the start that one aim in constructing this model theory was to clarify, in isolation, the separate concepts that interact with each other in strategic contexts, and that are the component parts of the complex concepts used to describe those contexts. In keeping with this motivation, I will first look at a pure and simple abstract version of belief revision theory, for a single agent in a single possible

that players' strategy choices are made independently: if player one had chosen differently, it could not have influenced the choice of player two. But this assumption of causal independence has no consequences about the evidential relevance of information about player one's choice for the beliefs that a third party might rationally have about player two. (Brian Skyrms (1992, pp. 147–8) makes this point.)

[8] This characterization theorem is proved in Stalnaker (1994).

world, ignoring degrees of belief, and assuming nothing about the subject matter of the beliefs. After getting clear about the basic structure, I will say how to incorporate it into our models, with many agents, many possible worlds, and probability measures on both the prior and posterior belief states. The simple theory that I will sketch is a standard one that has been formulated in a number of essentially equivalent ways by different theorists.[9] Sometimes the theory is formulated syntactically, with prior and posterior belief states represented by sets of sentences of some formal language, but I will focus on a purely model theoretic formulation of the theory in which the agent's belief revision policy is represented by a set of possible worlds – the prior belief state – and a function taking each piece of potential new information into the conditional belief state that corresponds to the state that would be induced by receiving that information. Let B be the set representing the prior state, and let B′ be the set of all the possible worlds that are compatible with any new information that the agent could possibly receive. Then if ϕ is any proposition which is a subset of B′, B(ϕ) will be the set that represents the posterior belief state induced by information ϕ.

There are just four constraints that the standard belief revision theory imposes on this belief revision function:

1. For any ϕ, $B(\phi) \subseteq \phi$
2. If ϕ is nonempty, then B(ϕ) is nonempty
3. If $B \cap \phi$ is nonempty, then $B(\phi) = B \cap \phi$
4. If $B(\phi) \cap \psi$ is nonempty, then $B(\phi \& \psi) = B(\phi) \cap \psi$

The first condition is simply the requirement that the new information received is believed in the conditional state. The second is the requirement that consistent information results in a consistent conditional state. The third condition requires that belief change be conservative in the sense that one should not give up any beliefs unless the new information forces one to give something up: if ϕ is compatible with the prior beliefs, the conditional belief state will simply add ϕ to the prior beliefs. The fourth condition is a generalization of the conservative condition. Its effect is to require that if two pieces of information are received in succession, the second being compatible with the posterior state induced by the first, then the resulting change should be the same as if both pieces of information were received together.

Any belief revision function meeting these four conditions can be represented by an ordering of all the possible worlds, and any ordering of a set of possible worlds will determine a function meeting the four conditions. Let Q be any binary transitive and connected relation on a set B′. Then we can define B as the set of highest ranking members of B′, and for any subset ϕ of B′, we can define B(ϕ) as the set of highest ranking members of ϕ:

[9] The earliest formulation, so far as I know, of what has come to be called the AGM belief revision theory was given by William Harper (1975). For a general survey of the belief revision theory, see Gärdenfors (1988). Other important papers include Alchourón and Makinson (1982), Alchourón et al. (1985), Grove (1988), Makinson (1985) and Spohn (1987).

$$\mathrm{B}\,(\phi) = \{\mathrm{x} \in \phi \;\; : \;\; \mathrm{yQx} \;\; \text{for all } \mathrm{y} \in \phi\}$$

It is easy to show that this function will satisfy the four conditions. On the other hand, given any revision function meeting the four conditions, we can define a binary relation Q in terms of it as follows:

$$\mathrm{xQy} \;\; \text{if } \mathrm{y} \in \mathrm{B}\,(\{\mathrm{x}, \mathrm{y}\})\,.$$

It is easy to show, using the four conditions, that Q, defined this way, is transitive and connected, and that $\mathrm{B}(\phi) = \{\mathrm{x} \in \phi\text{: yQx for all y} \in \phi\}$. So the specification of such a Q relation is just an alternative formulation of the same revision theory.

Now to incorporate this belief revision theory into our models, we need to give each player such a belief revision policy in each possible world. This will be accomplished if we add to the model a binary relation Q for each player. We need just one such relation for each player, if we take our assumption that players know their own states of mind to apply to belief revision policies as well as to beliefs themselves. Since the belief revision policy is a feature of the agent's subjective state, it is reasonable to assume that in all possible worlds that are subjectively indistinguishable for a player, he has the same belief revision policies.

Subjective indistinguishability (which we defined as follows: $\mathrm{x} \approx_{\mathrm{i}} \mathrm{y}$ if and only if $\{\mathrm{z:xR_iz}\} = \{\mathrm{z:yR_iz}\}$) is an equivalence relation that partitions the space of all possible worlds for each player, and the player's belief revision function will be the same for each world in the equivalence class. (The equivalence class plays the role of B' in the simple belief revision structure.) What we need to add to the game model is a relation $\mathrm{Q_i}$ for each player that orders all the worlds within each equivalence class with respect to epistemic plausibility, with worlds compatible with what the player believes in the worlds in that class having maximum plausibility. So $\mathrm{Q_i}$ must meet the following three conditions:

(q1) $\mathrm{x} \approx_{\mathrm{i}} \mathrm{y}$, if and only if $\mathrm{xQ_iy}$ or $\mathrm{yQ_ix}$.
(q2) $\mathrm{Q_i}$ is transitive.
(q3) $\mathrm{xR_iy}$ if and only if $\mathrm{wQ_iy}$ for all w such that $\mathrm{w} \approx_{\mathrm{i}} \mathrm{x}$.

For any proposition ϕ, we can define the conditional belief state for player i in world x, $\mathrm{B_{i,x}}(\phi)$ (the posterior belief state that would be induced by learning ϕ),[10] in terms of $\mathrm{Q_i}$ as follows:

$$\mathrm{B_{i,x}}\,(\phi) = \{\mathrm{w} \in \phi : \text{for all } \mathrm{y} \in \phi \cap \{\mathrm{z} : \mathrm{z} \approx_{\mathrm{i}} \mathrm{x}\}\,, \mathrm{yQ_iw}\}\,.$$

[10] There is this difference between the conditional belief state $\mathrm{B_{i,x}}(\phi)$ and the posterior belief state that would actually result if the agent were in fact to learn that ϕ: if he were to learn that ϕ, he would believe that he *then* believed that ϕ, whereas in our static models, there is no representation of what the agent comes to believe in the different possible worlds at some later time. But the potential posterior belief states and the conditional belief states as defined do not differ with respect to any information represented in the model. In particular, the conditional and posterior belief states do not differ with respect to the agent's beliefs about his *prior* beliefs.

Once we have added to our models a relation Q_i for each player that meets these three conditions, the R relations become redundant, since they are definable in terms of Q.[11] For a more economical formulation of the theory, we drop the R_i′s when we add the Q_i′s, taking condition (q1) as above the new definition of subjective indistinguishability, and condition (q3) as the definition of R_i. Formulated this way, the models are now defined as follows:

A model is a structure $<W,a,<S_i,Q_i,P_i>_{i \in N}>$. W, a, P_i, and S_i are as before; Each Q_i is a binary reflexive transitive relation on W meeting in addition the following condition: any two worlds that are Q_i related (in either direction) to a third world are Q_i related (in at least one direction) to each other. One can then prove that each R_i, defined as above, is serial, transitive, and euclidean. So our new models incorporate and refine models of the simpler kind.

To summarize, the new structure we have added to our models expresses exactly the following two assumptions:

1. In each possible world each player has a belief revision policy that conforms to the conditions of the simple AGM belief revision theory sketched above, where (for player i and world x) the set B is $\{y{:}xR_iy\}$, and the set B′ is $\{y{:}\ y \approx_i x\}$
2. In each world, each player has a correct belief about what his own belief revision policy is.

Each player's belief revision structure determines a ranking of all possible worlds with respect to the player's degree of epistemic success or failure in that world. In some worlds, the player has only true beliefs; in others, he makes an error, but not as serious an error as he makes in still other possible worlds. Suppose I am fifth on a standby waiting list for a seat on a plane. I learn that there is only one unclaimed seat, and as a result I feel certain that I will not get on the plane. I believe that the person at the top of the list will certainly take the seat, and if she does not, then I am certain that the second in line will take it, and so on. Now suppose in fact that my beliefs are mistaken: the person at the top of the list turns the seat down, and the next person takes it. Then my initial beliefs were in error, but not as seriously as they would be if I were to get the seat. If number two gets the seat, then I was making a simple first degree error, while if I get the seat, I was making a fourth degree error.

It will be useful to define, recursively, a sequence of propositions that distinguish the possible worlds in which a player's beliefs are in error to different degrees:

[11] The work done by Q is to rank the worlds incompatible with prior beliefs; it does not distinguish between worlds compatible with prior beliefs – they are ranked together at the top of the ordering determined by Q. So Q encodes the information about what the prior beliefs are – that is why R becomes redundant. A model with both Q and R relations would specify the prior belief sets in two ways. Condition (q3) is the requirement that the two specifications yield the same results.

Here is a simple abstract example, just to illustrate the structure: suppose there are just three possible worlds, x y and z, that are subjectively indistinguishable in those worlds to player i. Suppose {x} is the set of worlds compatible with i′s beliefs in x, y, and z, which is to say that the R relation is the following set: {<x,x>,<y,x>,<z,x>}. Suppose further that y has priority over z, which is to say if i were to learn the proposition {y,z}, his posterior or conditional belief state would be {y}. In other words, the Q relation is the following set: {<x,x>,<y,x>,<z,x>,<y,y>,<z,y>,<z,z>}.

$E^1{}_i$ is the proposition that player i has at least some false belief – makes at least a simple first degree error.

$E_i^1 = \{x \in W\text{: for some y such that } y \approx_i x, \text{ not } yQ_ix\} (= \{x \in W\text{: not } xR_ix\})$

$E^{k+1}{}_i$ is the proposition that player i makes at least a k + 1 degree error:

$$E_i^{k+1} = \left\{x \in E_i^k \ : \ \text{for some } y \in E_i^k \text{ such that } y \approx_i x, \text{ not } yQ_ix\right\}.$$

The belief revision structure provides for epistemic distinctions between propositions that are all believed with probability one. Even though each of two propositions has maximum degree of belief, one may be believed *more robustly* than the other in the sense that the agent is more disposed to continue believing it in response to new information. Suppose, to take a fanciful example, there are three presidential candidates, George, a Republican from Texas, Bill, a Democrat from Arkansas, and Ross, an independent from Texas. Suppose an agent believes, with probability one, that George will win. She also believes, with probability one, that a Texan will win and that a major party candidate will win, since these follow, given her other beliefs, from the proposition that George will win. But one of these two weaker beliefs may be more robust than the other. Suppose the agent is disposed, on learning that George lost, to conclude that Bill must then be the winner. In this case, the belief that a major party candidate will win is more robust than the belief that a Texan will win.

The belief revision structure is purely qualitative, but the measure functions that were already a part of the models provide a measure of the partial beliefs for conditional as well as for prior belief states. The Q relations, like the R relations, deliver the sets of possible worlds relative to which degrees of belief are defined. The partial beliefs for conditional belief state, like those for the prior states, are given by relativizing the measure function to the relevant set of possible worlds. Just as player i's partial beliefs in possible world x are given by relativizing the measure to the set $B_{i,x} = \{y{:}xR_iy\}$, so the partial beliefs in the conditional belief state for player i, world x and condition ϕ is given by relativizing the measure to the set $B_{i,x}(\phi) = \{y \in \phi\text{: for all } z \in \phi \text{ such that } z \approx_i x, zQ_iy)$.

So with the help of the belief revision function we can define *conditional* probability functions for each player in each world:

$$P_{i,x}\,(\phi/\psi) = P_i\,(\phi \cap B_{i,x}\,(\psi))\,/P_i(B_{i,x}\,(\psi)$$

In the case where the condition ψ is compatible with i′s prior beliefs – where $P_{i,x}(\psi) > 0$ – this will coincide with conditional probability as ordinarily defined. (This is ensured by the conservative condition on the belief revision function.) But this definition extends the conditional probability functions for player x in world i to any condition compatible with the set of worlds that are subjectively indistinguishable for x in i.[12]

[12] These extended probability functions are equivalent to *lexicographic probability systems.* See Blume et al. (1991a, b) for an axiomatic treatment of lexicographic probability in the context of decision theory and game theory. These papers discuss a concept equivalent to the one defined below that I am calling perfect rationality.

The belief revision theory, and the extended probability functions give us the resources to introduce a refinement of the concept of rationality. Say that an action is *perfectly rational* if it not only maximizes expected utility, but also satisfies a tie-breaking procedure that requires that certain *conditional* expected utilities be maximized as well. The idea is that in cases where two or more actions maximize expected utility, the agent should consider, in choosing between them, how he should act if he learned he was in error about something. And if two actions are still tied, the tie-breaking procedure is iterated – the agent considers how he should act if he learned that he were making an error of a higher degree. Here is a sequence of definitions leading to a definition of perfect rationality.

Given the extended conditional probability functions, the definition of conditional expected utility is straightforward:

$$\mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}/\phi\right) = \sum_{e \in \mathsf{C}_{-\mathrm{i}}} P_{i,x}\left([e]/\phi\right) \times u_i\left((s,e)\right)$$

Second, we define, recursively, a sequence of sets of strategies that maximize expected utility, and also satisfy the succession of tie-breaking rules:

$$\begin{aligned}
\mathrm{r}^{0}_{\mathrm{i,x}} &= \mathrm{r}_{\mathrm{i,x}}\left(\text{that is, } \left\{\mathrm{s} \in \mathrm{C_i} : \mathrm{eu}_{\mathrm{i,x}}(\mathrm{s}) \geq \mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}'\right) \text{ for all } \mathrm{s}' \in \mathrm{C_i}\right\}\right)\\
\mathrm{r}^{\mathrm{k+1}}_{\mathrm{i,x}} &= \left\{\mathrm{s} \in \mathrm{r}^{\mathrm{k}}_{\mathrm{i,x}} : \mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}/\mathrm{E}^{\mathrm{k+1}}\right) \geq \mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}'/\mathrm{E}^{\mathrm{k+1}}\right) \text{ for all } \mathrm{s}' \in \mathrm{r}^{\mathrm{k}}_{\mathrm{i,x}}\right\}.\\
\mathrm{r}^{+}_{\mathrm{i,x}} &= \cap \mathrm{r}^{\mathrm{k}}_{\mathrm{i,x}} \text{ for all k such that } \mathrm{E}^{\mathrm{k}} \cap \{\mathrm{y} : \mathrm{x} \approx_{\mathrm{i}} \mathrm{y}\} \text{ is nonempty.}
\end{aligned}$$

The set $\mathrm{r}^{+}_{\mathrm{i,x}}$ is the set of strategies that are perfectly rational for player i in world x. So the proposition that player i is perfectly rational is defined as follows:

$$\mathrm{A}^{+}_{\mathrm{i}} = \left\{\mathrm{x} \in \mathrm{W} : \mathrm{S_i}(\mathrm{x}) \in \mathrm{r}^{+}_{\mathrm{i,x}}\right\}.$$

I want to emphasize that this refinement is defined wholly within individual decision theory. The belief revision theory that we have imported into our models is a general, abstract structure, as appropriate for a single agent facing a decision problem to which the actions of other agents are irrelevant as it is for a situation in which there are multiple agents. It is sometimes said that while states with probability 0 are

I don't want to suggest that this is the only way of combining the AGM belief revision structure with probabilities. For a very different kind of theory, see Mongin (1994). In this construction, probabilities are nonadditive, and are used to represent the belief revision structure, rather than to supplement it as in the models I have defined. I don't think the central result in Mongin (1994) (that the same belief revision structure that I am using is in a sense equivalent to a nonadditive, and so non-Bayesian, probability conception of prior belief) conflicts with, or presents a problem for, the way I have defined extended probability functions: the probability numbers just mean different things in the two constructions.

relevant in game theory, they are irrelevant to individual decision making,[13] but I see no reason to make this distinction. There is as much or as little reason to take account, in one's deliberation, of the possibility that nature may surprise one as there is to take account of the possibility that one may be fooled by one's fellow creatures.

Perfect rationality is a concept of individual decision theory, but in the game model context this concept may be used to give a model theoretic definition of a refinement of rationalizability. Say that a strategy of a game Γ is *perfectly rationalizable* if and only if the strategy is played in some model of Γ in which the players have common belief that they all are perfectly rational. As with ordinary correlated rationalizability, one can use a simple algorithm to pick out the relevant class of strategies, and prove a characterization theorem that states that the model theoretic and algorithmic definitions determine the same class of strategies. Here is the theorem:

Strategies that survive the elimination of all weakly dominated strategies followed by the iterated elimination of strictly dominated strategies are all and only those that are realized in a model in which players have common belief that all are perfectly rational.[14]

Before going on to discuss knowledge, let me give two examples of games to illustrate the concepts of perfect rationality and perfect rationalizability.

First, consider the following very simple game: Alice can take a dollar for herself alone, ending the game, or instead leave the decision up to Bert, who can either decide whether the two players get a dollar each, or whether neither gets anything. Figure 42.1 represents the strategic form of this game.

Both strategies for both players are rationalizable, but only Tt is perfectly rationalizable. If Alice is certain that Bert will play t, then either of her strategies would maximize expected utility. But only choice T will ensure that utility is maximized also on the condition that her belief about Bert's choice is mistaken. Similarly, Bert may be certain that Alice won't give him the chance to choose, but if he has to commit himself to a strategy in advance, then if he is perfectly rational, he will opt for the choice that would maximize expected utility if he did get a chance to choose.

[13] For example, Fudenberg and Tirole (1992) make the following remark about the relation between game theory and decision theory: 'Games and decisions differ in one key respect: probability-0 events are both exogenous and irrelevant in decision problems, whereas what *would* happen if a player played differently in a game is both important and endogenously determined'.

To the extent that this is true, it seems to me an accident of the way the contrasting theories are formulated, and to have no basis in any difference in the phenomena that the theories are about.

[14] The proof of this theorem, and others stated without proof in this paper, are available from the author. The argument is a variation of the proof of the characterization theorem for simple (correlated) rationalizability given in Stalnaker (1994). See Dekel and Fudenberg (1990) for justification of the same solution concept in terms of different conditions that involve perturbations of the payoffs.

I originally thought that the set of strategies picked out by this concept of perfect rationalizability coincided, in the case of two person games, with perfect rationalizability as defined by Bernheim (1984), but Pierpaolo Battigalli pointed out to me that Bernheim's concept is stronger.

Fig. 42.1

		BERT	
		t	l
ALICE	T	0 1	0 1
	L	1 1	0 0

Fig. 42.2

		BERT	
		t	l
	T	2	2
ALICE	LT	1	3
	LL	1	0

Second, consider the following pure common interest game, where the only problem is one of coordination. It is also a perfect information game. One might think that coordination is no problem in a perfect information game, but this example shows that this is not necessarily true.

Alice can decide that each player gets two dollars, ending the game, or can leave the decision to Bert, who may decide that each player get one dollar, or may give the decision back to Alice. This time, Alice must decide whether each player gets three dollars, or neither gets anything. Figure 42.2 represents the strategic form of this game.

Now suppose Bert believes, with probability one, that Alice will choose T; what should he do? This depends on what he thinks Alice would do on the hypothesis that his belief about her is mistaken. Suppose that, if he were to be surprised by Alice choosing L on the first move, he would conclude that, contrary to what he previously believed, she is irrational, and is more likely to choose L on her second choice as well. Given these belief revision policies, only choice t is perfectly rational for him. But why should Alice choose T? Suppose she is sure that Bert will choose t, which as we have just seen, is the only perfectly rational choice for him to make if his beliefs about Alice are as we have described. Then Alice's only rational choice is T. So it might be that Alice and Bert both know each others' beliefs about each other, and are both perfectly rational, but they still fail to coordinate on the optimal outcome for both. Of course nothing in the game requires that Bert and Alice should have these beliefs and belief revision policies, but the game is compatible with them, and with the assumption that both Bert and Alice are perfectly rational.

Now one might be inclined to question whether Bert really believes that Alice is fully rational, since he believes she would choose L on her second move, if she got a second move, and this choice, being strictly dominated, would be irrational. Perhaps if Bert believed that Alice was actually disposed to choose L on her second move, then he wouldn't believe she was fully rational, but it is not suggested that he believes this. Suppose we divide Alice's strategy T into two strategies, TT and TL, that differ only in Alice's counterfactual dispositions: the two strategies are 'T, and I *would* choose T again on the second move if I were faced with that choice', and 'T, but I *would* choose L on the second move if I were faced with that choice'. One might argue that only TT, of these two, could be fully rational, but we may suppose that Bert believes, with probability one, that Alice will choose TT, and not TL. But were he to learn that he is wrong – that she did not choose TT (since she did not choose T on the first move) he would conclude that she instead chooses LL. To think there is something incoherent about this combination of beliefs and belief revision policy is to confuse epistemic with causal counterfactuals – it would be like thinking that because I believe that if Shakespeare hadn't written Hamlet, it would have never been written by anyone, I must therefore be disposed to conclude that Hamlet was never written, were I to learn that Shakespeare was in fact not its author.

Knowledge

As has often been noted, rationalizability is a very weak constraint on strategy choice, and perfect rationalizability is only slightly more restrictive. Would it make any difference if we assumed, not just common *belief* in rationality, or perfect rationality, but common *knowledge* as well? Whether it makes a difference, and what difference it makes, will depend on how knowledge is analyzed, and on what is assumed about the relation between knowledge and belief. I will consider a certain analysis of knowledge with roots in the philosophical literature about the definition of knowledge, an analysis that can be made precise with the resources of the belief revision structure that we have built into our models. But before getting to that analysis, I want to make some general remarks about the relation between knowledge and belief.

Whatever the details of one's analysis of knowledge and belief, it is clear that the central difference between the two concepts is that the first, unlike the second, can apply only when the agent is in fact correct in what he believes: the claim that i knows that ϕ, in contrast with the claim that i believes that ϕ, entails that ϕ, is true. Everyone knows that knowledge is different from belief – even from the extreme of belief, probability one – in this way, but sometimes it is suggested that this difference does not matter for the purposes of decision theory, since the rationality of a decision is independent of whether the beliefs on which it is based are in fact correct. It is *expected* utility, not the value of the actual payoff that I receive in the end, that is relevant to the explanation and evaluation of my actions, and expected utility cannot be influenced by facts about the actual world that do not affect my beliefs. But as soon as we start looking at one person's beliefs and knowledge about another's beliefs and knowledge, the difference between the two notions begins to matter. The assumption that Alice believes (with probability one) that Bert believes (with probability one) that the cat ate the canary tells us nothing about what Alice believes about the cat and the canary themselves. But if we assume instead that Alice *knows* that Bert *knows* that the cat ate the canary, it follows, not only that the cat in fact ate the canary, but that Alice knows it, and therefore believes it as well.

Since knowledge and belief have different properties, a concept that conflates them will have properties that are appropriate for neither of the two concepts taken separately. Because belief is a subjective concept, it is reasonable to assume, as we have, that agents have introspective access to what they believe, and to what they do not believe. But if we switch from belief to knowledge, an external condition on the cognitive state is imposed, and because of this the assumption of introspective access is no longer tenable, even for logically omniscient perfect reasoners whose mental states are accessible to them. Suppose Alice believes, with complete conviction and with good reason that the cat ate the canary, but is, through no fault of her own, factually mistaken. She *believes,* let us suppose, that she knows that the cat ate the canary, but her belief that she knows it cannot be correct. Obviously, no amount of introspection into the state of her own mind will reveal to her the fact that she lacks this knowledge. If we conflate knowledge and belief, assuming in general that i knows that ϕ if and only if i's degree of belief for ϕ is one, then we get a concept that combines the introspective properties appropriate only to the internal, subjective concept of belief with the success properties appropriate only to an external concept that makes claims about the objective world. The result is a concept of knowledge that rests on equivocation.

The result of this equivocation is a concept of knowledge with the familiar partition structure, the structure often assumed in discussions by economists and theoretical computer scientists about common knowledge, and this simple and elegant structure has led to many interesting results.[15] But the assumption that knowledge and common knowledge have this structure is the assumption that there can be no such thing as false belief, that while ignorance is possible, error is not. And since there is no false belief, there can be no disagreement, no surprises, and no coherent counterfactual reasoning.[16]

[15] Most notably, Robert Aumann's important and influential result on the impossibility of agreeing to disagree, and subsequent variations on it all depend on the partition structure, which requires the identification of knowledge with belief. See Aumann (1976) and Bacharach (1985). The initial result is striking, but perhaps slightly less striking when one recognizes that the assumption that there is no disagreement is implicitly a premise of the argument.

[16] If one were to add to the models we have defined the assumption that the R relation is reflexive, and so (given the other assumptions) is an equivalence relation, the result would be that the three relations, $\mathrm{R_i}$, $\mathrm{Q_i}$, and $\approx_{\mathrm{i}}$, would all collapse into one. There would be no room for belief revision,

It is sometimes suggested that if one were to analyze knowledge simply as true belief, then the result would be a concept of knowledge with this partition structure, but this is not correct. The conjunctive concept, true belief, will *never* determine a partition structure unless it is assumed that it is necessary that *all* beliefs are true, in which case the conjunctive concept would be redundant. For suppose there might be a false belief – that it might be that some person i believed that ϕ, but was mistaken. Then it is false that i truly believes that ϕ, and so if true belief satisfied the conditions of the partition structure, it would follows that i truly believes that he does not truly believe that ϕ, from which (since he believes ϕ) he could infer that ϕ is false. The point is that to assume negative introspection for true belief is to assume that a believer can distinguish, introspecrively, her true beliefs from her false beliefs, which implies (at least if she is consistent) that she won't have any false beliefs.

While it can never be reasonable to equate knowledge and belief in general, we can specify the contingent conditions under which knowledge and belief will coincide. If we assume about a particular situation that *as a matter of fact,* a person has no false beliefs, then (and only then) can we conclude that in that situation, knowledge and belief coincide. To get this conclusion, we need to make no assumptions about knowledge beyond the minimal one that knowledge implies true belief. The assumption we need to make is that full belief is a state that is subjectively indistinguishable from knowledge: that fully believing that ϕ is the same as fully believing that one knows that ϕ.

If we make the idealizing assumption about a particular game situation being modeled that no one has any false beliefs, and that it is common belief that no one has any false beliefs, then we can have the benefits of the identification of knowledge and belief without the pernicious consequences that come from equivocating between them. What we cannot and need not assume is that it is a *necessary* truth – true in all possible worlds in the model – that no one has any false beliefs. Even if players *actually* have only true beliefs, there will inevitably be *counterfactual* possible worlds in the model in which players have false beliefs. These counterfactual possible worlds must be there to represent the causal possibilities that define the structure of the game, and to represent the belief revision policies of the players. If we assumed that it was a necessary truth that there was no false belief, then it would be impossible for one player to believe that a second player was rational in any model for any game in which irrational options are available to the second player.

In terms of this idealizing assumption about knowledge and belief, we can define a refinement of rationalizability, which I have called *strong rationalizability*. Here

since it would be assumed that no one had a belief that could be revised. Intuitively, the assumption would be that it is a necessary truth that all players are Cartesian skeptics: they have no probability-one beliefs about anything except necessary truths and facts about their own states of mind. This assumption is not compatible with belief that another player is rational, unless it is assumed that it is a necessary truth that the player is rational.

is the model theoretic definition: for any game Γ a strategy profile is strongly rationalizable if and only if it is realized in a model in which there is no error, common belief that all players are rational, and common belief that there is no error. The set of strategy profiles characterized by this condition can also be given an algorithmic definition, using an iterated elimination procedure intermediate between the elimination of strictly dominated and of weakly dominated strategies.[17] We can also define a further refinement, *strong perfect rationalizability*: just substitute 'perfect rationality' for 'rationality' in the condition defining strong rationalizability. A minor variation of the algorithm will pick out the set of strategy profiles characterized by these conditions.

Knowledge and belief coincide on this demanding idealization, but suppose we want to consider the more general case in which a person may know some things about the world, even while being mistaken about others. How should knowledge be analyzed? The conception of knowledge that I will propose for consideration is a simple version of what has been called, in the philosophical literature about the analysis of knowledge, the *defeasibility analysis*. The intuitive idea behind this account is that 'if a person has knowledge, then that person's justification must be sufficiently strong that it is not capable of being *defeated* by evidence that he does not possess' (Pappas and Swain 1978). According to this idea, if evidence that is unavailable to you would give you reason to give up a belief that you have, then your belief rests in part on your ignorance of that evidence, and so even if that belief is true, it will not count as knowledge.

We can make this idea precise by exploiting the belief revision structure sketched above, and the notion of robustness that allowed us to make epistemic distinctions between propositions believed with probability one. The analysis is simple: i knows that ϕ if and only if i believes that ϕ (with probability one), *and that belief is robust with respect to the truth*. That is, i knows that ϕ in a possible world x if and only if ϕ receives probability one from i in x, and also receives probability one in every conditional belief state for which the condition is true in x. More precisely, the proposition that i knows that ϕ is the set $\{x \in W\text{: for all } \psi \text{ such that } x \in \psi,\ B_{i,x}(\psi) \subseteq \phi\}$.

Let me illustrate the idea with the example discussed above of the presidential candidates. Recall that there are three candidates, George, Bill and Ross, and that the subject believes, with probability one, that George will win. As a result she also believes with probability one that a Texan will win, and that a major party candidate will win. But the belief that a major party candidate will win is more robust than the belief that a Texan will win, since our subject is disposed, should she learn that George did not win, to infer that the winner was Bill. Now suppose, to everyone's surprise, Ross wins. Then even though our subject's belief that a Texan would win turned out to be true, it does not seem reasonable to say that she *knew* that a Texan would win, since she was right only by luck. Had she known more (that George

[17] The algorithm, which eliminates iteratively profiles rather than strategies, is given in Stalnaker (1994), and it is also proved there that the set of strategies picked out by this algorithm is characterized by the class of models meeting the model theoretic condition.

would lose), then that information would have undercut her belief. On the other hand, if Bill turns out to be the winner, then it would not be unreasonable to say that she knew that a major party candidate would win, since in this case her belief did not depend on her belief that it was George rather than Bill that would win.

The defeasibility conception of knowledge can be given a much simpler definition in terms of the belief revision structure. It can be shown that the definition given above is equivalent to the following: the proposition *i knows that* ϕ is the set $\{x: \{y:xQ_iy\} \subseteq \phi \}$. This exactly parallels the definition of the proposition that i believes that ϕ: $\{x: \{y:xR_iy\} \subseteq \phi\}$. On the defeasibility analysis, the relations that define the belief revision structure are exactly the same as the relations of epistemic accessibility in the standard semantics for epistemic logic.[18] And common knowledge (the infinite conjunction, everyone knows that ϕ, everyone knows that everyone knows that ϕ, ...) exactly parallels common belief: the proposition *there is common knowledge that* ϕ is $\{x:\{y:xQ^*y\} \subseteq \phi\}$, where Q^* is the transitive closure of the Q_i relations.

The defeasibility analysis provides us with two new model theoretic conditions that can be used to define solution concepts: first, the condition that there is common knowledge of rationality; second, the condition that there is common knowledge of perfect rationality. The conditions are stronger (respectively) than the conditions we have used to characterize rationalizability and perfect rationalizability, but weaker than the conditions that characterize the concepts I have called strong rationalizability and strong perfect rationalizability. That is, the class of models in which there is common belief in (perfect) rationality properly includes the class in which there is common knowledge, in the defeasibility sense, of (perfect) rationality, which in turn properly includes the class in which there is no error, common belief that there is no error, and common belief in (perfect) rationality. So the defeasibility analysis gives us two distinctive model theoretic solution concepts, but surprisingly, the sets of strategy profiles characterized by these new model theoretic conditions are the same as those characterized, in one case, by the weaker condition, and in the other case by the stronger condition. That is, the following two claims are theorems:

1. *Any strategy realized in a model in which there is common belief in (simple) rationality is also realized in a model in which there is common knowledge (in the defeasibility sense) of rationality.*
2. *Any strategy profile realized in a model in which there is common knowledge of perfect rationality is also realized in a model meeting in addition the stronger condition that there is common belief that no one has a false belief.*[19]

[18] The modal logic for the knowledge operators in a language that was interpreted relative to this semantic structure would be S4.3. This is the logic characterized by the class of Kripke models in which the accessibility relation is transitive, reflexive, and weakly connected (if xQ_iy and xQ_iz, then either yQ_iz or zQ_iy). The logic of common knowledge would be S4.

[19] Each theorem claims that any strategy that is realized in a model of one kind is also realized in a model that meets more restrictive conditions. In each case the proof is given by showing how to modify a model meeting the weaker conditions so that it also meets the more restrictive conditions.

Backward Induction

To illustrate how some of this apparatus might be deployed to help clarify the role in strategic arguments of assumptions about knowledge, belief and counterfactual reasoning, I will conclude by looking at a puzzle about backward induction reasoning, focusing on one notorious example: the finite iterated prisoners' dilemma. The backward induction argument purports to show that if there is common belief, or perhaps common knowledge, that both players are rational, then both players will defect every time, from the beginning. Obviously rational players will defect on the last move, and since they know this on the next to last move, they will defect then as well, and so on back through the game. This kind of argument is widely thought to be paradoxical, but there is little agreement about what the paradox consists in. Some say that the argument is fallacious, others that it shows an incoherence in the assumption of common knowledge of rationality, and still others that it reveals a self-referential paradox akin to semantic paradoxes such as the liar. The model theoretic apparatus we have been discussing gives us the resources to make precise the theses that alternative versions of the argument purport to prove, and to assess the validity of the arguments. Some versions are clearly fallacious, but others, as I will show, are valid.

The intuitive backward induction argument applies directly to games in extensive form, whereas our game models are models of static strategic form games.[20] But any extensive form game has a unique strategic form, and proofs based on the idea of the intuitive backward induction argument can be used to establish claims about the strategic form of the game. A backward induction argument is best seen as an argument by mathematical induction about a class of games that is closed with respect to the subgame relation – in the case at hand, the class of iterated prisoners' dilemmas of length n for any natural number n.

The conclusions of the backward induction arguments are conditional theses: if certain conditions obtain, then players will choose strategies that result in defection every time. The conditions assumed will correspond to the constraints on models that we have used to characterize various solution concepts, so the theses in question will be claims that only strategy profiles that result in defection every time will satisfy the conditions defining some solution concept. If, for example, the conditions are that there is common belief in rationality, then the thesis would be that only strategies that result in defection every time are rationalizable. It is clear that a backward induction argument for this thesis must be fallacious since many

[20] Although in this paper we have considered only static games, it is a straightforward matter to enrich the models by adding a temporal dimension to the possible worlds, assuming that players have belief states and perform actions at different times, actually revising their beliefs in the course of the playing of the game in accordance with a belief revision policy of the kind we have supposed. Questions about the relationship between the normal and extensive forms of games, and about the relations between different extensive-form games with the same normal form can be made precise in the model theory, and answered.

cooperative strategies are rationalizable. Pettit and Sugden (1989) have given a nice diagnosis of the fallacy in this version of the argument. But what if we make the stronger assumption that there is common *knowledge* of rationality, or of perfect rationality? Suppose, first, that we make the idealizing assumption necessary for identifying knowledge with belief: that there is no error and common belief that there is no error, and common belief that both players are rational. Are all *strongly* rationalizable strategy pairs in the iterated prisoners' dilemma pairs that result in defection every time? In this case the answer is positive, and the theorem that states this conclusion is proved by a backward induction argument.

To prove this backward induction theorem, we must first prove a lemma that is a general claim about multi-stage games – a class of games that includes iterated games. First, some notation and terminology: let Γ be any game that can be represented as a multi-stage game with observed action (a game that can be divided into stages where at each stage all players move simultaneously, and all players know the result of all previous moves). Let $\Gamma^{\#}$ be any subgame – any game that begins at the start of some later stage of Γ. For any strategy profile c of Γ that determines a path through the subgame $\Gamma^{\#}$, let $c^{\#}$ be the profile for $\Gamma^{\#}$ that is determined by c, and let $C^{\#}$ be the set of all strategy profiles of Γ that determine a path through $\Gamma^{\#}$. By 'an SR model', I will mean a model in which there is (in the actual world of the model) no error, common belief that there is no error, and common belief that all players are rational. Now we can state the multi-stage game lemma:

If profile c is strongly rationalizable in Γ, and if c determines a path through $\Gamma^{\#}$, then $c^{\#}$ is strongly rationalizable in $\Gamma^{\#}$.

This is proved by constructing a model for $\Gamma^{\#}$ in terms of a model for Γ, and showing that if the original model is an SR model, so is the new one. Let M be any SR model for Γ in which c is played in the actual world of the model. Let $\Gamma^{\#}$ be any subgame that contains the path determined by c. We define a model $M^{\#}$ for $\Gamma^{\#}$ in terms of M as follows: $W^{\#} = \{x \in W{:}S(x) \in C^{\#}\}$. The $Q_i{}^{\#\prime}$s and $P_i{}^{\#\prime}$s are simply the restrictions of the $Q_i{}'$s and $P_i{}'$s to $W^{\#}$. The $S_i{}^{\#\prime}$s are defined so that for each $x \in W^{\#}$, $S^{\#}$ (x) is the profile for the game $\Gamma^{\#}$ that is determined by the profile S(x). (That is, if $S(x) = e$, then $S^{\#}(x) = e^{\#}$.)

To see that $M^{\#}$ is an SR model for $\Gamma^{\#}$, note first that if there is no error and common belief that there is no error in the original model, then this will also hold for the model of the subgame: if $\{x{:}aR^{*}x\} \subseteq \{x{:}xR_ix \text{ for all } i\}$, then $\{x{:}aR^{\#*}x\} \subseteq \{x{:}xR_i{}^{\#}x \text{ for all } i\}$. This is clear, since $\{x{:}aR^{\#*}x\} \subseteq \{x{:}aR^{*}x) \cap W^{\#}$, and $(x{:}xR_i{}^{\#}x \text{ for all } i) = \{x{:}xR_ix \text{ for all } i\} \cap W^{\#}$. Second, because of the fact that players know all previous moves at the beginning of each stage, they can make their strategy choices conditional on whether a subgame is reached. (More precisely, for any player i and pair of strategies s and s′ for i, that are compatible with $\Gamma^{\#}$ being reached, there is a strategy equivalent to this: s if $\Gamma^{\#}$ is reached, s′ if not.) This implies that for any world w, player i and subgame such that it is compatible with i′s beliefs that that subgame be reached, a strategy will be rational for i only if the strategy determined for the subgame is rational, conditional on the hypothesis that the subgame is reached. This ensures that rationality is preserved in all worlds when the model is modified. So $c^{\#}$ is strongly rationalizable in $\Gamma^{\#}$.

An analogous result about strong *perfect* rationalizability can be shown by essentially the same argument.

One further observation before turning to the backward induction theorem itself: for any game Γ, if profile c is compatible with common belief in (the actual world of) an SR model for Γ, then c itself is strongly rationalizable. It is obvious that if $S(x) = c$ and $aR^{*}x$, then the same model, with x rather than a as the actual world will be an SR model if the original model was.

Now the backward induction theorem:

Any strongly rationalizable strategy profile in a finite iterated prisoners' dilemma is one in which both players defect every time.

The proof is by induction on the size of the game. For the base case – the one shot PD – it is obvious that the theorem holds, since only defection is rational. Now assume that the theorem holds for games of length k. Let Γ be a game of length $k + 1$, and Γ^{-} be the corresponding iterated PD of length k. Let M be any SR model of Γ, and let c be any strategy profile that is compatible with common belief (that is, c is any profile for which there exists an x such that $S(x) = c$, and $aR^{*}x$). By the observation just made, c is strongly rationalizable, so by the multi-stage game lemma, c^{-} (the profile for Γ^{-} determined by c) is strongly rationalizable in Γ^{-}. But then by hypothesis of induction, c^{-} is a profile in which both players defect every time. So c (in game Γ) is a profile in which both players defect every time after the first move. But c is any profile compatible with common belief in the actual world of the model, so it follows that in the model M, it is common belief that both players will choose strategies that result in defection every time after the first move. Given these beliefs, any strategy for either player that begins with the cooperative move is strictly dominated, relative to that player's beliefs. So since the players are both rational, it follows that they choose a strategy that begins with defection, and so one that results in defection on every move.

Our theorem could obviously be generalized to cover some other games that have been prominent in discussions of backward induction such as the centipede game and (for strong *perfect* rationalizability) the chain store game. But it is not true, even in perfect information games, that the strong or strong and perfect rationalizability conditions are always sufficient to support backward induction reasoning. Recall the perfect information, pure coordination game discussed above in which Alice and Bert failed to coordinate on the backward induction equilibrium, even though the conditions for strong perfect rationalizability were satisfied. In that example, the strategy profile played was a perfect, but not subgame perfect, equilibrium. One can show in general that in perfect information games, all and only Nash equilibrium strategy profiles are strongly rationalizable (see Stalnaker (1994) for the proof).

As I noted at the end of the last section, it can be shown that the set of strongly and perfectly rationalizable strategy profiles is characterized also by the class of models in which there is common knowledge (in the defeasibility sense) of perfect rationality. So we can drop the strong idealizing assumption that there is no error, and still get the conclusion that if there is common knowledge (in the defeasibility sense) of perfect rationality, then players will choose strategies that result in defection every time.

Pettit and Sugden, in their discussion of the paradox of backward induction, grant that the argument is valid when it is common knowledge rather than common belief that is assumed (though they don't say why they think this, or what they are assuming about knowledge). But they suggest that there is nothing surprising or paradoxical about this, since the assumption of common *knowledge* of rationality is incompatible with the possibility of rational deliberation, and so is too strong to be interesting. Since knowledge logically implies truth, they argue, the argument shows that 'as a matter of logical necessity, both players *must* defect and presumably therefore that they know they must defect' (Pettit and Sugden 1989). But I think this remark rests on a confusion of epistemic with causal possibilities. There is no reason why I cannot both *know* that something is true, and also entertain the counterfactual possibility that it is false. It is of course inconsistent to suppose, counterfactually or otherwise, the conjunction of the claim that ϕ is false with the claim that I know that ϕ is true, but it is not inconsistent for me, knowing (in the actual world) that ϕ is true, to suppose, counterfactually, that ϕ is false. As Pettit and Sugden say, the connection between knowledge and truth is a matter of logical necessity, but that does not mean that if I know that I will defect, I therefore *must* defect, 'as a matter of logical necessity'. One might as well argue that lifelong bachelors are powerless to marry, since it is a matter of logical necessity that lifelong bachelors never marry.

The semantic connection between knowledge and truth is not, in any case, what is doing the work in this version of the backward induction argument: it is rather the assumption that the players believe in common that neither of them is in error about anything. We could drop the assumption that the players beliefs are all actually true, assuming not common knowledge of rationality, but only common belief in rationality and common belief that no one is in error about anything. This will suffice to validate the induction argument.

Notice that the common belief that there will not, in fact, be any surprises, does not imply the belief that there couldn't be any surprises. Alice might think as follows: 'Bert expects me to defect, and I will defect, but I could cooperate, and if I did, he would be surprised. Furthermore, I expect him to defect, but he could cooperate, and if he did, I would be surprised'. If these 'could's were epistemic or subjective, expressing uncertainty, then this soliloquy would make no sense, but it is unproblematic if they are counterfactual 'could's used to express Alice's beliefs about her and Bert's capacities. A rational person may know that she will not exercise certain of her options, since she may believe that it is not in her interest to do so.

It is neither legitimate nor required for the success of the backward induction argument to draw conclusions about what the players would believe or do under counterfactual conditions. In fact, consider the following 'tat for tit' strategy: defect on the first move, then on all subsequent moves, do what the other player did on the previous move, until the last move; defect unconditionally on the last move. Our backward induction argument does not exclude the possibility that the players should each adopt, in the actual world, this strategy, since this pair of strategies results in defection every time. This pair is indeed compatible with the conditions for strong and perfect rationalizability. Of course unless each player assigned a very low probability to the hypothesis that this was the other player's strategy, it would not be rational for him to adopt it, but he need not rule it out. Thus Pettit and Sugden are wrong when they say that the backward induction argument can work only if it is assumed that each player would maintain the beliefs necessary for common belief in rationality 'regardless of what the other does' (Pettit and Sugden 1989, p. 178). All that is required is the belief that the beliefs necessary for common knowledge of rationality will, in fact, be maintained, given what the players in fact plan to do. And this requirement need not be *assumed*: it is a consequence of what is assumed.

Conclusion

The aim in constructing this model theory was to get a framework in which to sharpen and clarify the concepts used both by rational agents in their deliberative and strategic reasoning and by theorists in their attempts to describe, predict and explain the behavior of such agents. The intention was, first, to get a framework that is rich in expressive resources, but weak in the claims that are presupposed or implicit in the theory, so that various hypotheses about the epistemic states and behavior of agents can be stated clearly and compared. Second, the intention was to have a framework in which concepts can be analyzed into their basic components, which can then be considered and clarified in isolation before being combined with each other. We want to be able to consider, for example, the logic of belief, individual utility maximization, belief revision, and causal-counterfactual structure separately, and then put them together to see how the separate components interact. The framework is designed to be extended, both by considering further specific substantive assumptions, for example, about the beliefs and belief revision policies of players, and by adding to the descriptive resources of the model theory additional structure that might be relevant to strategic reasoning or its evaluation, for example temporal structure for the representation of dynamic games, and resources for more explicit representation of counterfactual propositions. To illustrate some of the fruits of this approach we have stated some theorems that provide model theoretic characterizations of some solution concepts, and have looked closely at one familiar form of reasoning – backward induction – and at some conditions that are sufficient to validate this form of reasoning in certain games, and at conditions that are not sufficient. The focus has been on the concepts involved in two kinds of counterfactual reasoning whose interaction is essential to deliberation in strategic contexts, and to the evaluation of the decisions that result from such deliberation: reasoning about what the consequences would be of actions that are alternatives to the action chosen, and reasoning about how one would revise one's beliefs if one were to receive information that one expects not to receive. We can get clear about why people do what they do, and about what they ought to do, only by getting clear about the relevance of what they could have done, and might have learned, but did not.[21]

References

Adams, E. (1970). Subjunctive and indicative conditionals. *Foundations of Language, 6*, 89–94.

Alchourón, C., & Makinson, D. (1982). The logic of theory change: Contraction functions and their associated revision functions. *Theoria, 48*, 14–37.

Alchourón, C., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic, 50*, 510–530.

Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics, 4*, 1236–1239.

Bacharach, M. (1985). Some extensions of a claim of Aumann in an axiomatic model of knowledge. *Journal of Economic Theory, 37*, 167–190.

Bernheim, B. (1984). Rationalizable strategic behavior. *Econometrica, 52*, 1007–1028.

Blume, L., Brandenburger, A., & Dekel, E. (1991a). Lexicographic probabilities and choice under uncertainty. *Econometrica, 59*, 61–79.

Blume, L., Brandenburger, A., & Dekel, E. (1991b). Lexicographic probabilities and equilibrium refinements. *Econometrica, 59*, 81–98.

Dekel, E., & Fudenberg, D. (1990). Rational behavior with payoff uncertainty. *Journal of Economic Theory, 52*, 243–267.

Fudenberg, D., & Tirole, J. (1992). *Game theory*. Cambridge, MA: MIT Press.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge, MA: MIT Press.

Gibbard, A., & Harper, W. (1981). Counterfactuals and two kinds of expected utility. In C. Hooker et al. (Eds.), *Foundations and applications of decision theory*. Dordrecht/Boston: Reidel.

Grove, A. (1988). Two modelings for theory change. *Journal of Philosophical Logic, 17*, 157–170.

Harper, W. (1975). Rational belief change, popper functions and counterfactuals. *Synthese, 30*, 221–262.

Lewis, D. (1980). Causal decision theory. *Australasian Journal of Philosophy, 59*, 5–30.

Makinson, D. (1985). How to give it up: A survey of some formal aspects of the logic of theory change. *Synthese, 62*, 347–363.

Mongin, P. (1994). The logic of belief change and nonadditive probability. In D. Prawitz & D. Westerstahl (Eds.), *Logic and philosophy of science in Uppsala*. Dordrecht: Kluwer.

Pappas, G., & Swain, M. (1978). *Essays on knowledge and justification*. Ithaca: Cornell University Press.

Pearce, G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica, 52*, 1029–1050.

Pettit, P., & Sugden, R. (1989). The backward induction paradox. *Journal of Philosophy, 86*, 169–182.

Skyrms, B. (1982). Causal decision theory. *Journal of Philosophy, 79*, 695–711.

Skyrms, B. (1992). *The dynamics of rational deliberation*. Cambridge, MA: Harvard University Press.

Spohn, W. (1987). Ordinal conditional functions: A dynamic theory of epistemic states. In W. Harper & B. Skyrms (Eds.), *Causation in decision, belief change and statistics* (Vol. 2, pp. 105–134). Dordrecht: Reidel.

Stalnaker, R. (1994). On the evaluation of solution concepts. *Theory and Decision, 37*, 49–73.

[21] I would like to thank Pierpaolo Battigalli, Yannis Delmas, Drew Fudenberg, Philippe Mongin, Hyun Song Shin, Brian Skyrms, and an anonymous referee for helpful comments on several earlier versions of this paper.