PAUL WEIRICH

INTRODUCTION: INTERACTIVE EPISTEMOLOGY

My daughter finishes school at 3:00 p.m. She knows I know this. So she expects me to pick her up then. An agent's knowledge of others agents' knowledge influences her expectations and therefore her behavior. Interactive epistemology studies agents' knowledge of agents' knowledge (their interactive knowledge), the events that shape its structure, and its effect on the agents' behavior (for example, their agreeing to disagree). Taken broadly, it covers agents' beliefs and degrees of belief, as well as their knowledge. The titles of landmark articles by Robert Aumann (1999a, 1999b) canonize the term interactive epistemology.

This issue presents current research in interactive epistemology. It covers common knowledge, backward induction, arguments for common prior probabilities, interactive knowledge's influence on the step from degrees of belief to betting rates, and the combination of dynamic epistemic logic and game theory. Before previewing the articles, this introduction provides some background.

Suppose that every member of a group knows that p, knows that every member knows that p, knows that every member knows that every member knows that p, and so on. Then the group has common knowledge that p. A proposition's public announcement may make the proposition common knowledge. Common knowledge is a key idea of interactive epistemology. Peter Vanderschraaf and Giacomo Sillari (2009) review its various characterizations, its origin, its role in inferences, and its significance in game theory.

Mutual knowledge that p obtains in a group just in case all in the group know that p. Common knowledge entails (1) mutual knowledge that p, (2) mutual knowledge that mutual knowledge that p, and so on, up an infinite hierarchy of mutual knowledge. To couch the definition of common knowledge in compact notation, notation for mutual knowledge helps. Let $(EK)p$ stand for everyone knows that p. Then let $(EK)^2p$ abbreviate $(EK)(EK)p$, let $(EK)^np$ abbreviate a string of n (EK)'s followed by p, and let $(EK)^\infty p$ abbreviate for all n, $(EK)^np$. A group has common knowledge that p if and only if $(EK)^\infty p$.

This familiar characterization of common knowledge leaves open arrangement of quantifiers. Does "all know that all know that p" mean that each knows that all know that p, or that each knows that each knows that p? Letting K_xp abbreviate x knows that p, does $(EK)^2p$ stand for $\forall xK_x\forall yK_yp$ or $\forall x\forall yK_xK_yp$? The usual disambiguation takes the second interpretation. Hence, in a group, a member's

Episteme 8.3 (2011): 201–208
DOI: 10.3366/epi.2011.0017

www.eupjournals.com/epi

knowing that all know that *p* is equivalent to the member's knowing a conjunction in which each conjunct expresses a member of the group's knowledge that *p*. The member may know the conjunction without knowing that it covers every member of the group.

David Lewis introduced common knowledge to explain why people observe conventions. For realism, he characterized common knowledge in terms of reasons to believe rather than knowledge.

> Let us say that it is *common knowledge* in a population *P* that ____ if and only if some state of affairs *A* holds such that:
>
> (1) Everyone in *P* has reason to believe that *A* holds.
> (2) *A* indicates to everyone in *P* that everyone in *P* has reason to believe that *A* holds.
> (3) *A* indicates to everyone in *P* that ____. (1969, 56)

According to this definition, common knowledge that *p* exists in a population when the members of the population have an infinite hierarchy of reasons for mutual beliefs, that is, beliefs throughout the population. In a population mutual beliefs ascend only a few steps up the hierarchy of reasons. Some theorists say that justified mutual beliefs further up the hierarchy are implicit or tacit. Furthermore, mutual beliefs held may not amount to knowledge. They may be false despite being justified by reasons in the hierarchy. Only in ideal cases do beliefs constitute knowledge and ascend the whole hierarchy of reasons. Robin Cubitt and Robert Sugden (2003) formalize the distinctive features of Lewis's account of common knowledge.

Aumann (1976) characterizes common knowledge using epistemic logic, and using set theory rather than predicate logic. He defines communal possibility for a population in terms of epistemic possibility for members of the population. At a world ω, $M(\omega)$ stands for the set of communally possible worlds. A world ω' is communally possible if and only if within the population a path of epistemic accessibility goes from ω to ω'. More precisely, $\omega' \in M(\omega)$ if and only if for some n-tuple of members of the population (not necessarily n distinct members), at ω for the first member a world is possible such that at it for the second member a world is possible such that at it for the third member a world is possible … such that at it for the nth member ω' is possible. Suppose that agents are ideal and aware of other agents' epistemic possibilities. Aumann proves that in a world ω, a proposition *p* is common knowledge (in the standard sense) if and only if *p* is true in each element of $M(\omega)$. Using communal possibility, agents may conduct inferences about common knowledge without working through an infinite hierarchy of mutual knowledge.

Inferences involving common knowledge depend on common knowledge's characterization and the inference's objective. In a familiar story illustrating the relevance of common knowledge to action, two women, Alice and Betty, strangers to each other, are traveling in a compartment of a train. Each woman's face bears the grime of travel. She sees that the other woman's face is dirty but does not see

https://doi.org/10.3366/epi.2011.0017 Published online by Cambridge University Press

that her own face is dirty. Their compartment tickets issue turns for the lavatory. As they know, Alice's turn is first and Betty's turn is second. Each knows that the other cleans her face if and only if she knows it is dirty. Neither woman knows her face is dirty, so neither woman cleans her face.

The train's conductor enters the compartment to collect tickets and, to be helpful, mentions that someone has a dirty face. Each woman already knew this, but now knows that each knows this, knows that each knows this, and so on. The two women acquire common knowledge that at least one has a dirty face.

After the conductor's announcement, Alice forgoes her turn to use the lavatory. She thinks that Betty's dirty face prompted the conductor's announcement. Betty observes that Alice does not clean her face. She infers that Alice does not clean her face because Alice sees that Betty's face is dirty. So Betty uses her turn to wash her face. Common knowledge leads Betty to the conclusion that her face is dirty.

Inferences involving common knowledge are crucial in games of strategy. Toward the end of a game of Tic-Tac-Toe, suppose a player knows that unless she puts an X in the middle of the bottom row, her opponent will put an O there and win the game. Her prediction of her opponent's behavior uses her knowledge that he knows he wins if he puts an O in the middle of the bottom row. Such knowledge of her opponent's knowledge directs her placement of an X. A method of strategic reasoning, backward induction, predicts the move of the last player to move, then the move of the second to last player to move, and so on until the first player makes a move using the predictions about subsequent moves. Common knowledge supports backward induction in sequential games, in which players take turns making moves.

The vexing backward induction paradox arises from the observation that backward induction's predictions assume that players are rational. If a player were to deviate from these predictions, the deviation would furnish evidence that the player is not rational. An opponent's subsequent move should then not rest on predictions that assume the player's rationality. Backward induction's assumptions appear to be unjustified when it predicts moves following irrational moves. Cristina Bicchieri (1989) and Philip Pettit and Robert Sugden (1989) present the paradox. Robert Aumann (1995), Ken Binmore (1996), and Robert Stalnaker (1996, 1998) debate its resolution.

In a simultaneous-move game, players make moves at the same time, and so each moves without observation of the other players' moves. If the players have common knowledge of their game's payoff matrix and their rationality, strict domination may simplify their choices. For a player, one strategy strictly dominates another strategy if and only if the first strategy is better than the second strategy, no matter what the other players do. A rational player eliminates a strictly dominated strategy. Other players, knowing the game's payoff matrix and the player's rationality, know that he eliminates that strategy. Hence they remove it from the payoff matrix, and each considers whether for her any strategy strictly

https://doi.org/10.3366/epi.2011.0017 Published online by Cambridge University Press

dominates another strategy in the reduced matrix. If so, she eliminates it, and the others, knowing that she knows the payoff matrix and is rational, know that she eliminates it. Hence they remove it from the payoff matrix. This process continues until the payoff matrix, after reductions, no longer contains for any player a strictly dominated strategy. Players choose among the strategies that remain after iterated elimination of strictly dominated strategies.

Besides supporting iterated elimination of strictly dominated strategies, the players' common knowledge of their game and their rationality assists the case for their realization of a solution to their game, that is, a profile of strategies, one for each player, such that each strategy is rational given the profile. Without that common knowledge, a rational player may not perform her part of a solution because she lacks confidence that the other players will do their parts. Adam Brandenburger (1992) explains how common knowledge of various features of a game support familiar criteria for a solution.

Aumann and Brandenburger (1995) observe that utility maximization and knowledge of the strategy profile realized entail realization of a Nash equilibrium, a profile of strategies such that each is a best response to the others. Only a Nash equilibrium is such that, given knowledge of its realization, each of its strategies maximizes utility. Although this result relies on mutual knowledge, not common knowledge, of the profile realized, common knowledge of the game and the players' rationality explains mutual knowledge of the profile realized. The explanation requires a few glosses on the players' common knowledge.

As a background assumption, a reader of a textbook's presentation of a game supposes that all the players know what he knows about the game. Because the reader knows every relevant feature of the game, the assumption entails that the players know every relevant feature of the game too. What a player knows is a relevant feature of the game because it affects the player's behavior. The reader knows every relevant fact that any player knows. Hence, by assumption, the players all know every relevant fact that any player knows. This result (given standard epistemic logic) is enough to generate the players' common knowledge of any relevant fact that any player knows. In the idealized games a textbook presents, the players have common knowledge of a relevant proposition if at least one player knows the proposition.

In an ideal game, the assumption that players have common knowledge of all relevant features of their game replaces the weaker assumption that players have common knowledge of their payoff matrix. Some knowledge, such as knowledge of the game's rules, a player obtains directly. Other knowledge, such as the player's knowledge of her choice and its rationality, she infers during deliberations. A player's knowledge of the strategy profile realized comes through inference.

In an ideal game, the players know the strategy profile realized, whatever it is. For every profile, each player knows that the profile is realized if it is realized. This hypothetical knowledge covers every profile and not just the profiled realized.

Knowledge of the profile realized, whatever it is, comes from a player's knowledge of his own strategy, whatever it is, and knowledge of the other players' responses to his strategy. A player's knowledge of responses to his strategies I call prescience. Prescience explains knowledge of the profile realized and also knowledge of any profile hypothetically realized. Prescience is a product of players' common knowledge of their game and their rationality. Their common knowledge explains the knowledge of conditionals that constitutes prescience.

Imagine a game with the players Row and Column. Row may move Up or Down, and Column may move Left or Right. Suppose that if Row moves Up, Column's best response is Right. Knowledge of conditionals such as "If Up, then Right" arises from common knowledge. It comes from a player's knowledge of his own strategy and an opponent's adopting a best response to his strategy, and all this being common knowledge. If Row adopts Up, then he knows this. If he knows his strategy, then Column knows that he knows. If Column knows that, then Column knows his strategy. Hence Column adopts a best response. Column knows her response. Hence Row knows her response. Everything relevant that anyone knows is common knowledge. Hence it is common knowledge that if Row adopts Up, then Column adopts Right. By such reasoning, each player attains prescience of responses. Each knows for each strategy the other's response. Common knowledge explains prescience, knowledge of the profile realized, and so realization of a Nash equilibrium.

Part of a Nash equilibrium's explanation is a generalization of the principle of utility maximization. The generalization treats decision problems in which an agent's choice provides information about the state of the world that settles the consequences of his choice. It handles decision problems in games of strategy where a player's choice furnishes information about other players' choices. According to the generalization, a rational player adopts a strategy that maximizes utility given its adoption. The strategy is self-ratifying, and so not regretted as soon as adopted. To verify that a strategy is self-ratifying, a player assumes that she adopts the strategy. Then she assigns probabilities to other players' strategies under that assumption. The assumption grounds the probability assignment.

In a two-person game with a unique Nash equilibrium in mixed strategies, a pure strategy has the same expected payoff as a player's Nash strategy given that his opponent participates in the equilibrium. Why does the player adopt his Nash strategy? Prescience and ratification furnish reasons. Adopting a pure strategy gives an agent evidence that his opponent adopts a best response and so a deviation from her Nash strategy. Given her deviation, his pure strategy does not maximize utility. It is not self-ratifying.

Weirich (2010, sect. 6.6) shows how prescience arising from common knowledge supports a Nash equilibrium in an ideal game with a unique Nash equilibrium, provided that rational agents comply with the principle of ratification. It does this for the mixed extension of Matching Pennies using a proof that generalizes to every two-person game with a unique Nash equilibrium.

https://doi.org/10.3366/epi.2011.0017 Published online by Cambridge University Press

These brief remarks about common knowledge, backward induction, and Nash equilibrium introduce interactive epistemology. The essays that follow advance the field. They display its richness and subtlety.

Zachary Ernst's essay, "What Is Common Knowledge?" examines assumptions of standard characterizations of common knowledge. Because of these assumptions, familiar models of agents' knowledge fail to distinguish between full-blown common knowledge and merely finite levels of interactive (or mutual) knowledge. Ernst suggests revising accounts of common knowledge to respond more faithfully to the motivation for Lewis's definition of common knowledge. This revision requires attention to cognitive limits, the value of conventions for cognitively limited agents, and human reliance on heuristics rather than complicated inferences. The accompanying account of rationality lowers the bar for choices made under constraints the real world imposes.

In "Logic in a Social Setting," Johan van Benthem treats information-driven social action. He describes the effect that backward induction's account of solutions to extensive (or sequential) games has on the logic of those games. Dynamic epistemic logic explains the origin of the players' common knowledge of their rationality. A generalization of backward induction uses agents' beliefs, in particular their rankings of the game's possible outcomes according to plausibility. Other generalizations revise beliefs in light of agents' previous moves in a game, revise preferences attributed to agents in light of their previous moves, and use information about agents' computational limits and methods of belief revision. Logic's job is to abstract uniformities from the complicated models that the generalizations introduce for the sake of realism. Accounts of social action, which include reasoning, elucidate logic itself. The combination of logic and game theory generalized is a theory of play.

In a game with a sequence of stages, backward induction instructs a player calculating how an opponent will move at a stage to ignore the opponent's moves at earlier stages. Whether this is rational depends on the assumptions governing the game, including the interpretation of counterfactuals about stages that the players do not reach. Ken Binmore, in "Interpreting Knowledge in the Backward Induction Problem," considers whether common knowledge of rationality implies backward induction in finite games of perfect information, in which a player at a stage knows the moves made at earlier stages. Binmore argues that if knowledge is interpreted as commitment, then the implication holds, whereas if knowledge is interpreted as certitude, then the implication does not hold. He illustrates his points with several simple sequential games.

Ashton Sperry-Taylor's essay, "Bounded Rationality in the Centipede Game," formulates a theory of rationality for agents with cognitive limits. The essay shows that agents complying with the normative theory behave in the Centipede Game, a game in which payoffs increase if agents forego incentives for immediate payoffs, just as do people in experimental studies of the game. The normative theory relaxes game theory's usual idealizations about agents' cognitive powers, on which

https://doi.org/10.3366/epi.2011.0017 Published online by Cambridge University Press

backward induction's solution to the Centipede Game rests, and takes account of a person's limited ability to look ahead and calculate an opponent's responses to moves. Using bounded rationality to account for observed behavior in the Centipede Game works better than does appealing to social norms, and it enriches plausible explanations that other theories provide.

Luc Bovens and Wlodek Rabinowicz, in "Bets on Hats," present a case in which a group of rational agents with the same goals appear to be open to a Dutch book, a system of bets that guarantees a loss. They show how strategic considerations save the group from exploitation. In the case, individual rationality in pursuit of a common goal leads to collective rationality. Their essay formulates for groups a restricted version of Bas van Fraassen's principle of reflection. The new principle asks members of a group to adjust present probabilities in light of members' future probability assignments, but only given certain conditions concerning the members' prior probabilities and future evidence. Their essay also analyzes the logic of Dutch book arguments and assesses the interpretation of degrees of belief as betting rates.

Itzhak Gilboa, in "Why the Empty Shells Were Not Fired," points out a problem with an initially appealing argument that agents have a common prior probability assignment before they acquire information and update their probability assignments. The argument treats agents in their uninformed state as "empty shells" ready to receive information. Gilboa argues that these uninformed agents, if all are epistemically alike and ignorant of their identities, cannot learn their identities. Updating from a common prior therefore cannot account for agents' self-knowledge.

These six first-rate essays are a representative sample of the burgeoning field of interactive epistemology. They invite additional research on the fascinating ideas that characterize the field. Hearty thanks to Alvin Goldman for encouraging the collection and directing its production.

REFERENCES

Aumann, Robert. 1976. "Agreeing to Disagree." *Annals of Statistics* 4: 1236–9.

Aumann, Robert. 1995. "Backward Induction and Common Knowledge of Rationality." *Games and Economic Behavior* 8: 6–19.

Aumann, Robert. 1999a. "Interactive Epistemology I: Knowledge." *International Journal of Game Theory* 28: 263–300.

Aumann, Robert. 1999b. "Interactive Epistemology II: Probability." *International Journal of Game Theory* 28: 301–14.

Aumann, Robert and Adam Brandenburger. 1995. "Epistemic Conditions for Nash Equilibrium." *Econometrica* 63: 1161–80.

Bicchieri, Cristina. 1989. "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge." *Erkenntnis* 30: 69–85.

Binmore, Ken. 1996. "A Note on Backward Induction." *Games and Economic Behavior* 17: 135–7.

Brandenburger, Adam. 1992. "Knowledge and Equilibrium in Games." *Journal of Economic Perspectives* 6: 83–101.

Cubitt, Robin and Robert Sugden. 2003. "Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory." *Economics and Philosophy* 19: 175–210.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Pettit, Philip and Robert Sugden. 1989. "The Backward Induction Paradox." *Journal of Philosophy* 86: 169–82.

Stalnaker, Robert. 1996. "Knowledge, Belief and Counterfactual Reasoning in Games." *Economics and Philosophy* 12: 133–63.

Stalnaker, Robert. 1998. "Belief Revision in Games: Forward and Backward Induction." *Mathematical Social Sciences* 36: 31–56.

Vanderschraaf, Peter and Giacomo Sillari. 2009. "Common Knowledge." In E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/spr2009/entries/common-knowledge/

Weirich, Paul. 2010. *Collective Rationality: Equilibrium in Cooperative Games*. New York: Oxford University Press.