Unformatted final version. Published in *Nous* 53, 2 (2019): 491-510

THE SORITICAL CENTIPEDE

Nathan Ballantyne, Brian Fiala and Terry Horgan

Fordham University, Dartmouth College and University of Arizona

Two philosophical questions arise about rationality in centipede games that are logically prior to attempts to apply the formal tools of game theory to this topic. First, given that the players have common knowledge of mutual rationality and common knowledge that they are each motivated solely to maximize their own profits, is there a backwards-induction argument that (i) employs only familiar non-technical concepts about rationality, (ii) leads to the conclusion that the first player is rationally obligated to end the game at the first step, (iii) is deductively valid, (iv) employs premises all of which are prima facie highly plausible, and (v) is prima facie sound (in virtue of features (iii) and (iv))? Second, if there is such an argument, then is it actually sound, or is it instead defective somehow despite being prima facie sound? Addressing these two questions is our project. We present a backwards-induction argument that is prima facie sound; we argue that it is an instance of the notorious sorites paradox, and hence that the concepts of rational obligatoriness and rational permissibility are vague; and we briefly address the potential consequences of all this for the foundations of game theory and decision theory.

Centipede games are frequently discussed in game theory.[1] Here is how a typical game works: a pile of 101 one-dollar coins is placed on a table between two players of the game, player A and player B. The players take turns making moves in the game, with player A going first. At any potential stage of the game prior to stage 100, the player whose turn it is has two choices: either take one coin from the pile—in which case the coin now becomes that player's to keep, and the game continues—or take two coins from the pile—in which case those two coins now both become that player's to keep and the game stops (with neither player receiving any of the remaining coins in the pile). If the game reaches stage 100, then it ends at that stage regardless whether player B, the chooser at that stage, takes one coin or two.

In the extensive literature on centipede games in game theory, there is controversy about what constitutes rational behavior, under various background assumptions that might be made about the two players. Suppose, for example (as is frequently done), that there is common knowledge of mutual rationality and common knowledge that each player is motivated solely by the goal of maximizing her/his own profit in the game. That is, both players are rational and are solely egocentrically profit-motivated, they both know this, they both know that they know this, and so forth.

Some game theorists maintain that under these assumptions, rationality requires each player to have a strategy profile that exhibits subgame perfect Nash equilibrium—which means that the player's strategy profile is such that for any potential stage S, the profile dictates an action for S that would belong to a Nash equilibrium for a centipede game in which S is the initial stage. (A Nash equilibrium is a pair of strategies, one for each player, such that neither player could do better by unilaterally following a different strategy.) This entails, by backwards-induction reasoning, that rationality requires the first player to "defect" on the first play of the game, thereby ending it (in the above example, by taking two coins on the first play).

Other game theorists maintain instead that under the given assumptions, rationality only requires each player to have a profile of strategies that are "rationalizable"—where rationalizable strategies are those that survive a process of iterated elimination of strictly dominated strategies. (Eliminate any strategy that is worse for one player than some alternative strategy for that player regardless of what the other player does; then eliminate, from among the remaining strategies, any strategy that is worse for one player than some alternative strategy for that player regardless of what the other player does; and so on.) Many such profiles satisfy this less exacting criterion, including strategy profiles in which the players "cooperate" (in the above example, by taking only one coin) for much of the game.

Despite the substantial literature applying the formal tools of game theory to issues about rationality in centipede games, certain philosophical questions arise about this topic that have received scant attention and really are logically prior to attempts to apply those formal tools to the topic. Two such questions are the following. First, given that the players have common knowledge of mutual rationality and common knowledge that they are each motivated solely to maximize their own profits, is there a backwards-induction argument that (i) employs only familiar non-technical concepts about rationality (e.g., the concept of rational impermissibility and the concept of rational obligatoriness), (ii) leads to the conclusion that the first player is rationally obligated to end the game at the first step, (iii) is deductively valid, (iv) employs premises all of which are prima facie highly plausible, and (v) is prima facie sound (in virtue of features (iii) and (iv))? Second, if there is such an argument, then is it actually sound, or is it instead defective somehow despite being prima facie sound? These questions are of considerable philosophical interest in their own right, and they become all the more important because of the unresolved controversies in game theory about how best to formally model practical rationality vis-à-vis centipede games.

Addressing these two questions is our project here. First, we will formulate a backwards-induction argument that we claim exhibits the lately-mentioned features: it is deductively valid and prima facie sound, it deploys only familiar pre-theoretic ideas about rationality, and it leads to the conclusion that rationality requires the first player to end the game on the first round. Second, we will argue that despite being prima facie sound, this argument is actually unsound; it is an instance of the notorious sorites paradox. Third, since vagueness is the source of the sorites paradox, we will urge the further conclusion that the notion of practical rationality is itself vague—and essentially so. Fourth and finally, we will briefly discuss some apparent implications of all this for the foundations of game theory and decision theory.

1. Preliminaries: Dynamic vs. Static Centipede Games

One way to try constructing a backwards-induction argument for the conclusion that the first player is rationally required to defect on the first move is to reason counterfactually and dynamically: ask what *would* be a rationally required move if the game *were* at the final possible stage; then, in light of one's answer to that question, ask what would be a rationally required move if the game were at the next-to-last possible stage; and so on, successively backwards to stage 1. Here one is considering a so-called "extensive form" version of the game.

Another approach, however, instead focuses on a *static* version—a version in what is often called "normal form" or "strategic form." Here each player is required to choose a strategy right at the start (without knowing the other player's choice), and must stick to that choice. A strategy for a given player is a specification, for each potential stage of the game at which it is the player's turn to act, whether to cooperate or defect at that stage. (This is sometimes called a 'strategy profile, with 'strategy' being used in a more fine-grained way for a decision about what to do at any single stage.) There needn't be any actual game-playing at all in a static game; rather, after each player chooses a strategy, the two strategies can be revealed and the players can then be paid immediately whatever is coming to them given the respective strategies they have chosen.

Various complications arise regarding backwards-induction reasoning as applied to the extensive-form version that do not arise when such reasoning is applied to the static version—complications that generate controversy among game theorists about whether such reasoning, vis-à-vis extensive-form versions, is problematic (or perhaps outright unsound) for reasons *distinct* from possible soriticality. It is quite plausible, for instance, that for at least one potential stage of the dynamic game (e.g., the very last one), that stage could only be reached via a lapse in rationality by one of the players. Thus, even if at the start of the dynamic game there is common knowledge of mutual rationality, it is doubtful that this common knowledge *would* be present in all possible stages of the game. Robert Stalnaker (1998) puts the point well. He writes:

> Even if I think you *know* what I am going to do, I can consider how I think you would react if I did something that you and I both know I will not do, and my answers to such counterfactual questions will be relevant to assessing the rationality of what I *am* going to do. (p. 31)

This leads Stalnaker to say the following about the relevance, for practical rationality in game situations, of common knowledge of (or common belief in) mutual rationality:

> [W]hat can be said about how rational players should respond to surprising information? Very little, I will argue. That is, assumptions about rationality, and about common belief in rationality, put no substantive constraints on how an agent does or should revise beliefs in response to surprising information. We can, however, say a great deal about the consequences for action of various assumptions about belief revision policies, and of the assumptions about agents' beliefs and common beliefs about the belief revision policies of others. (p. 32)

The apparent import regarding centipede games is that counterfactually dynamic backwards-induction reasoning concerning extensive-form games is only sound given certain further assumptions about commonly known (or commonly believed-to-obtain) belief-revision policies of the respective players. Issues now arise about what such assumptions should be, and about how plausible they are. For instance, Sobel (1993) reconstructs backwards induction reasoning in a way that explicitly invokes these two assumptions:

> *Resilient rationality*: each player is disposed to act rationally at each possible node that the game can reach, including nodes that will certainly never be reached in actual play.
>
> *Robustness*: each player's beliefs in the players' future rationality would be kept come what may, whatever evidence of irrationality would by then transpire concerning past performance of the players.

These assumptions are very strong, arguably implausibly strong—a fact that threatens to severely diminish the interest of the resulting backwards-induction arguments regarding extensive-form centipede games.[2] On the other hand, Rabinowicz (1998) shows that for a class of extensive games that includes centipede games—*BI-terminating games*, as he calls them,

> [I]t is enough to make rationality assumptions concerning *actual* play; stipulations about counterfactual developments are not needed. Essentially, it is enough to assume that the first player (i) makes a rational move, (ii) believes that his successor (if there is to be one) will make a rational move, (iii) believes that his successor will have a corresponding belief about *his* successor, etc. (pp. 97-98)

He adds this: "The relevant 'if' is interpreted as weakly as possible—as a material implication" (p. 106). Yet, as Rabinowicz himself acknowledges, the conclusion of his proof is only that the first player *will* end the game on the first move, not that this is *rationally required*. Echoing the first of the two above-quoted passages from Stalnaker (1998), Rabinowicz writes:

> There is a troublesome feature of our proof…. We have seen that, in a BI-terminating game under conditions of forward rationality, the first player to move will chose the backward-induction move m. But what are his *reasons* for performing m rather than m′? We do not know enough to give a definite answer to this question. But if m is to be rational, the first player must hold appropriate beliefs about what would have happened in the continuation of the game if he had acted otherwise. (p. 112)

Thus, Rabinowicz's proof of this result does not constitute or entail a specification of the first player's *rationale* for ending the game on the first move—which leaves it unclear whether or not backwards-induction reasoning vis-à-vis a dynamic centipede game can itself constitute a credible such reason without importing problematically strong assumptions like resilience and robustness.[3,4]

Fortunately, for present purposes it will suffice to focus on the static version of the centipede game. This is so for several interconnected reasons. First, since (as we will maintain) a backwards-induction argument can be constructed for the static version that exhibits features (i)-(v) mentioned in penultimate paragraph of the introductory section above, this fact itself has significant philosophical interest. In particular, it provides strong evidence for the contention that the static backwards-induction argument is actually an instance of the sorites paradox—which entails that this argument is unsound in the same way(s) that other sorites arguments are unsound. Second, if indeed the static backwards-induction argument is soritical (as we will maintain it is), then this fact by itself establishes that the key notion deployed in the argument (viz., *rational impermissibility*) is sorites-susceptible—and hence is vague. Third, if this is so then the vagueness of the notion of rational impermissibility thereby infects backwards-induction reasoning vis-à-vis extensive-form (i.e., dynamic) centipede games as well, and thus renders such reasoning soritical too—since in both cases the reasoning invokes the putative, stepwise-backward, "spread" of the category of rational impermissibility (as putatively applying to take-only-one-coin moves) from one stage to the next-preceding stage. This means that however well or badly backwards-induction arguments vis-à-vis extensive-form centipede games might fare in other respects, these arguments too—like backwards-induction arguments vis-à-vis strategic-form centipede games—are sorites arguments and hence and unsound in whatever way(s) sorites arguments in general are unsound.

So the ensuing discussion will be about *static* centipede games in which the players have common knowledge of mutual rationality and also have common knowledge that each player is motivated solely by the goal of maximizing her or his own personal gain. It bears emphasis, however, that the discussion will extend, *mutatis mutandis*, to extensive-form centipede games too—and will be orthogonal to ongoing disputes in game theory about whether or not backwards induction in extensive-form centipede games is already objectionable for other reasons.

2. A Non-Technical Backwards-Induction Argument

We are seeking to formulate in pre-theoretic non-technical terms a deductive argument in favor of the claim that in static centipede games in which the two common-knowledge assumptions are in force, choosing the strategy of "defecting" on the first move of the game is rationally obligatory.

For specificity, we will focus on the static version of the particular centipede game we described in the first paragraph of the paper. For this game, the following 102 strategies are available, with the odd-numbered ones available to player A and the even-numbered ones available to player B:

1. Player A takes two coins at stage 1

2. Player B takes two coins at stage 2

3. Player A takes one coin at stage 1 and takes two coins at stage 3

4. Player B takes one coin at stage 2 and takes two coins at stage 4

5. Player A takes one coin at stages 1 and 3 and takes two coins at stage 5

. Player B takes one coin at stages 2 and 4 and takes two coins at stage 6

.

.

.

99. Player A takes one coin at stages 1, 3, …, 97 and takes two coins at stage 99

100. Player B takes one coin at stages 2, 4, …, 98 and takes two coins at stage 100

101. Player A takes one coin at stages 1, 3, …, 99

102. Player B takes one coin at stages 2, 4, …, 100

In a dynamic version of the game, it would be possible for each player to adopt such a policy at the beginning of the game, and for the players to retain their respective policies throughout the game and act accordingly—although, unless player A adopted strategy 101 and player B adopted either strategy 100 or strategy 102, one of the players would not complete her or his chosen strategy because the other would end the game before that could happen. In the static version we are discussing, each player gets paid whatever that player would have received in a dynamic version in which they both resolutely stuck strategies they actually chose.[5]

As we will formulate the backwards-inductive argument, it deploys both the notion of rational impermissibility and the correlative notion of rational obligatoriness. It also employs, as premises, backwardly-successive instances of the following backwards-induction schemas for the centipede game:

$\mathcal{C}_{A,B}(\alpha)$ [where $\alpha = 101$ or 99 or 97 or … or 7 or 5]:

If strategies 101 and 99 and … and α are rationally impermissible for player A, then strategy α-1 is rationally impermissible for player B.

$\mathcal{C}_{B,A}(\beta)$ [where $\beta = 100$ or 98 or 96 or… or 8 or 6 or 4]:

If strategies 102 and 100 and … and β are rationally impermissible for player B, then strategy β-1 is rationally impermissible for player A.

(Note that schema $\mathcal{C}_{A,B}(\alpha)$ has an initial limit-case instance involving just one "conjunct" in its antecedent—viz., the instance $\mathcal{C}_{A,B}(101)$.) The two schemas encode the key idea that the argument will deploy, viz. this:

> For any strategy S_i that a player P might adopt, except for the strategy of taking two coins on player P's first turn, if every higher-numbered strategy that the other player might adopt is rationally impermissible for the other player, then strategy S_i is rationally impermissible for player P.

(If the other player is not permitted to adopt any higher-numbered strategy than strategy S_i, then the other player must adopt some lower-numbered strategy than S_i—which guarantees that P's net profit under strategy S_i is less than the net profit that P might perhaps gain by instead adopting strategy S_{i-2}.) Utilizing this idea, the argument goes as follows:

1. Strategy 102 is rationally impermissible for player B. (Premise)
2. Strategy 101 is rationally impermissible for player A. (Premise)
3. If strategy 101 is rationally impermissible for player A, then strategy 100 is rationally impermissible for player B. (Premise: $\mathcal{C}_{A,B}(101)$)
4. Strategy 100 is rationally impermissible for player B. (2, 3, MP)
5. Strategies 102 and 100 are rationally impermissible for player B. (1, 4, Conj)
6. If strategies 102 and 100 are rationally impermissible for player B, then strategy 99 is rationally impermissible for player A. (Premise: $\mathcal{C}_{B,A}(100)$)
7. Strategy 99 is rationally impermissible for player A. (5, 6, MP)

8. Strategies 101 and 99 are rationally impermissible for player A. (2, 7, Conj)

.

.

.

293. Strategies 102 and 100 and … and 4 are rationally impermissible for player B. (287, 292, Conj)

294. If strategies 102 and 100 and … and 4 are rationally impermissible for player B, then strategy 3 is rationally impermissible for player A. (Premise: ($\mathcal{C}_{B,A}(4)$))

295. Strategy 3 is rationally impermissible for player A. (293, 294, MP)

296. Strategies 101 and 99 and … and 3 are rationally impermissible for player A. (290, 295, Conj)

297. If strategies 101 and 99 and … and 3 are rationally impermissible for player A, then strategy 1 is rationally obligatory for player A. (Premise)

298. Strategy 1 is rationally obligatory for player A. (296, 297, MP)

This formulation satisfies the desiderata laid out above. It is unquestionably valid, being articulable within standard propositional logic (with the instances of schemas $\mathcal{C}_{A,B}(\alpha)$ and $\mathcal{C}_{B,A}(\beta)$ all being ordinary material conditionals) and employing only the highly non-tendentious inference rules Modus Ponens and Conjunction. Premises 1 and 2 are clearly true, given that each player is rational and seeks to maximize his or her own profits. Premise 297 is clearly true, since rationality renders a specific strategy obligatory for a player if all of that player's other available strategies are rationally impermissible for that player. And, given the background assumptions of common knowledge of mutual rationality and common knowledge that each player is motivated solely by egocentric profit-maximization, the remaining premises—each of which is an instance of one or the other of the backwards-induction schemas $\mathcal{C}_{A,B}(\alpha)$ and $\mathcal{C}_{B,A}(\beta)$—all are prima facie highly plausible.

So the answer to the first question we posed at the outset is affirmative: there is indeed a presumptively sound, non-technically formulable, backwards-induction argument concluding that in (static) centipede games, the first player must choose the strategy that ends the game on the first move.

Nonetheless, such arguments are highly paradoxical from a common-sense point of view. After all, the players know full well that they both will do much better financially by both choosing strategies that keep the game going for a long while—even though it is puzzling just what to think about strategy-pairs that keep the game going until at or near the last possible stage. So the second question we initially posed now arises. Is the argument defective despite being presumptively sound, and if so then how?

3. Refuting the Argument

Although the argument as formulated above is *presumptively* sound and therefore deserves serious philosophical respect, we deny that it is *actually* sound. We come not to praise the non-technical backwards-induction argument, but to bury it. We claim that the argument is an instance of the notorious sorites paradox, and it is therefore fallacious in whatever way(s) other soritical arguments are fallacious. (We also claim that the argument cannot plausibly be refuted in any other way. In virtue of its presumptive soundness, either it is soritical or else it is actually sound.)

Consider, for example, the feature *heaphood*. The following principle (formulated as a schema) seems prima facie very plausible:

$\mathcal{H}$ If a pile of σ grains of sand is a heap, then a pile of σ-1 grains of sand is a heap.

Yet, if one embraces this principle along with, say, the premise that a pile of 10 million grains of sand is a heap, then one can construct a *presumptively* sound—albeit hugely paradoxical—sorites argument for the conclusion that, say, a pile of 20 grains of sand is a heap. One just invokes successive instances of principle H and successive applications of Modus Ponens, drawing successive conclusions about heaphood: a pile of 10 million minus one grain is a heap; a pile of 10 million minus two grains is a heap; …; a pile of 20 grains is a heap.

Virtually any vague concept is sorites-susceptible, in the sense that one can construct paradoxical sorites arguments deploying that concept. Moreover, the backwards-induction argument in Section 2 certainly has structural aspects like those exhibited in paradigmatic sorites arguments: repeated, stepwise applications of the same category to successive items in a sequence each member of which differs only slightly from its immediate neighbors—with the successive differences being uni-directional in some pertinent respect. (In the argument, the successively applied category is *being rationally impermissible*, and the successive items in the sequence are strategies 102 and 101, strategy 100, strategy 99, and so on. The successive instances of the backwards-induction schemas $\mathcal{C}_{A,B}(\alpha)$ and $\mathcal{C}_{B,A}(\beta)$ are the analogues of the successive instances, in a sorites argument concerning heaphood, of the successive instances of the schema $\boldsymbol{\mathcal{H}}$.) So the argument is at least an *eligible candidate* for being a sorites argument.

Is it one? Some might think that backwards-induction reasoning in centipede games is *obviously* soritical—and that such reasoning is therefore fallacious-qua-soritical whether or not it can be non-technically formulated in a way that frees it of other defects. Something close to this attitude is manifested by John Collins and Achille Varzi (2000). After describing a backwards-induction argument regarding a centipede game, they say, "We take the above story to imply that rationality predicates are, to some degree, vague" (p.3). This vagueness claim, which they treat as obvious, launches their argument in favor of their principal thesis—viz., that rationality predicates possess a form of vagueness that is "unsharpenable." (We return to this thesis in Appendix 3.) Our own epistemic phenomenology regarding backwards-induction arguments in centipede games is like that of Collins and Varzi: such arguments seem to us to be obviously soritical. Nonetheless, it is desirable to provide an *argument* for the vagueness of the concept of rational impermissibility—especially since, with the exception of Collins and Varzi (2000), the charge of soriticality virtually never surfaces in the literature on centipede games.

Three considerations tell strongly in favor of the vagueness of these rationality concepts, and in combination these considerations mutually reinforce one another evidentially. First, almost all concepts deployed or deployable by humans are vague to some extent—not only the concepts employed in the course of everyday life, but also most concepts in most branches of empirical science. The principal exceptions are in the formal sciences—pure mathematics, logic, set theory, and the like—and perhaps in some parts of theoretical physics (e.g., classical and quantum field theories). Therefore, there is already a strong *default assumption* that the correlative concepts of rational impermissibility and rational permissibility are vague to some extent. The burden of proof falls heavily on the shoulders of those who would deny that they are.[6]

Second, almost everyone has a deep and persistent intuition that backwards-induction reasoning regarding centipede games is highly paradoxical, and that the conclusion of such reasoning is blatantly false. The intuition applies to static centipede games too—the ones we have focused on here. This stubborn and widespread intuition cries out for explanation. All else equal, the best explanation will be one that adverts to people's semantic/conceptual competence with the notions of rational impermissibility and rational obligatoriness—as distinct from a proffered explanation that attributes the intuition to a stubborn, widespread performance error in the deployment of these concepts. And all else *is* equal here, since these are familiar concepts that people deploy all the time (albeit often implicitly) in everyday life. But because backwards-induction reasoning can be non-technically formulated in a way that renders it *presumptively* sound, a viable competence-based explanation of the pertinent intuition requires the hypothesis that the concepts of rational impermissibility and rational permissibility are vague. So the stubbornness and pervasiveness of the intuition together provide strong abductive support for this vagueness hypothesis.

Third, almost everyone, upon first pondering the matter, not only finds it deeply counterintuitive to claim that rationality requires the first player to end the game on the first move, but also finds it somewhat puzzling, and somewhat unclear, what would constitute a rationally optimal strategy to choose in a static centipede game. Although the very first steps in backwards-induction reasoning seem intuitively quite compelling, puzzlement and unclarity quickly start to set in as one contemplates successive backwards-induction steps beyond the very first one; correlatively, although iterating the backwards-induction steps very far seems clearly mistaken, one finds oneself lacking a firm, principled, basis for repudiating any *specific* step as mistaken. All this cries out for explanation, too. Once again, the best explanation will be a non-debunking explanation that adverts to people's semantic/conceptual competence with the notions of rational impermissibility and rational obligatoriness. And, if rational impermissibility is indeed vague with respect to centipede games, then one's competence *should* generate exactly such puzzlement and unclarity; likewise, it *should* leave one unable to classify any specific stage of the game as a stage for which a rationally optimal strategy will dictate defection rather than cooperation. For, according to the vagueness hypotheses, there is *no definite fact of the matter* about this very issue.[7] So people's sense of unclarity, about which specific stage of the static game, if any, is one for which an optimal strategy would dictate defecting, provides yet further abductive support for this vagueness hypothesis.

Considered individually, each of these considerations already constitutes significant evidence for the vagueness of rationality concepts. But when they are considered in tandem, their net evidential import is even stronger than the "sum of the parts." Once they are fed together into the hopper of wide reflective equilibrium, the upshot is a very strong case for the vagueness hypothesis.[8] And, given the vagueness hypothesis, the non-technically formulated backwards-induction argument concerning the static centipede game turns out to be fallacious. It is an instance of the infamous-yet-fallacious sorites paradox. Likewise, *mutatis mutandis*, for the dynamic version of the game—regardless of how well or badly backwards-induction reasoning in dynamic centipede games fares in other respects. Thus, the appropriate way to block the argument is to reject the principles $\mathcal{C}_{A,B}(\alpha)$ and $\mathcal{C}_{B,A}(\beta)$.[9]

4. Apparent consequences

The soriticality of backwards-induction reasoning regarding centipede games apparently has several important consequences, beyond undermining the reasoning itself. (We say 'apparently' in order to leave open the possibility of somehow resisting some or all of the apparent consequences we will describe.) First, it apparently undermines the applicability to centipede games of the notion of *expected utility*, thereby rendering inapplicable the normative principle of expected-utility maximization. To see this, suppose (for reductio) that under the usual common-knowledge assumptions, it is rationally permissible for a player P to assign a specific probability distribution to the various propositions of the form "The other player would execute strategy i if play were to continue long enough"; and suppose that P adopts some such probability distribution D. Then, assuming that P's utilities for the various possible outcomes are linear with the quantity of money P obtains, expected-utility maximization will require P to adopt a strategy that has, relative to D, a maximal expected utility for P. But this runs contrary to the fact that practical rationality is essentially vague in its application to static centipede games—a consequence of the soriticality of backwards-induction reasoning vis-à-vis these games. Hence, no such probability distribution is rationally permissible for P—which renders the notion of expected utility inapplicable in centipede games involving the usual common-knowledge assumptions.

Second, the soriticality of these backwards-induction arguments, and the consequent vagueness of practical-rationality notions, apparently means that practical rationality cannot be adequately defined as the obligatoriness of choosing an act or strategy with maximal expected utility, and practical rational permissibility cannot be adequately defined as the permissibility of choosing any of several acts or strategies all of which have the same, maximal, expected utility.[10] In centipede games involving the usual common-knowledge assumptions, practical rationality requires taking just one coin until very late in the game, and hence it is not rationally permissible to do otherwise—even though the notion of expected utility is inapplicable.

A third apparent consequence of the soriticality of these backwards-induction arguments is the need to acknowledge that the scope of standard game theory and standard decision theory is, to some extent, limited—because it does not include centipede games involving the usual common-knowledge assumptions. Let us note three different limited-scope positions.

Strong optimism asserts that the quantitative notions of subjective utility and subjective probability are applicable in the vast majority of decision/strategy problems, i.e., that in such problems these notions describe psychologically real features of actual human agents; it is only in certain unusual decision/strategy problems that these notions become inapplicable. On this view, expected-utility maximization normally *coincides* with practical rationality, even though the latter notion cannot be definitionally equated with the former one.

Moderate optimism asserts that mathematically precise models of rational decision-making are useful even if they are not literally applicable in the vast majority of decision/strategy problems. According to the moderate optimist, even if the notions of subjective utility and subjective probability normally do not describe psychologically real features of actual human agents, these notions often can be legitimately and fruitfully applied to construct *theoretical models* of rational decision-making and rational strategy-formation. The moderate optimist will claim that mathematical models deploying idealizing assumptions are a commonplace throughout science, and that game theory and decision theory are no worse off in this respect than any other branch of science. However, the moderate optimist will also concede that such models sometimes break down by engendering fallacious soritical reasoning—and that this is what happens in the case of centipede games.

Pessimism asserts that for most real-life decision/strategy problems, the mathematically precise concepts of game theory and decision theory neither describe psychologically real features of human agents nor provide illuminating theoretical models of rational decision-making and rational strategy-formation. Rather, says the pessimist, game theory and decision theory are only applicable in a quite limited domain, viz., decision/strategy problems in which it is psychologically realistic to suppose that an agent has quantitatively precise utilities for possible outcomes of the available acts or strategies, and has quantitatively precise subjective probabilities for the respective pertinent states of the world—e.g., gambling situations with known potential payoffs and known objective odds.[11]

A fourth apparent consequence of the soriticality of backwards-induction arguments concerning centipede games is the following. Game theory and decision theory are highly mathematical, and mathematical theorizing typically aspires to the same kind of precision that is manifested in *pure* mathematics—a precision that eschews vague concepts. This being so, there is a significant apparent tension between the vagueness of practical-rationality notions on one hand, and on the other hand the project of theorizing about practical rationality in a mathematically precise manner. This tension brings practitioners of decision theory and game theory, insofar as they regard themselves as explicating the ordinary pre-theoretic concept of practical rationality, face to face with Aristotle's famous remark about ethics, which seems no less applicable here: "Our account…will be adequate if it achieves such clarity as the subject-matter allows; for the same degree of precision is not to be expected in all discussions, any more than in all products of handicraft." (*Nicomachean Ethics*, Book 1, Chapter 3)

Appendix 1: On Common Knowledge of Mutual Rationality

We have argued that the backwards-induction reasoning set forth in Section 2 is soritical, even if one makes the standard assumptions (1) that there is common knowledge of mutual rationality and (2) that there is common knowledge that each player is motivated solely by maximizing her or his own financial gain. It might be objected, however, that our argument is unconvincing as long as assumption (1) is in play—and that therefore we should retreat to a weaker claim, viz., that our argument only applies to more realistic players rather than the idealized players of classical game theory.[12]

A natural way to motivate this objection would be to say that common knowledge of mutual rationality (for the static centipede game in question), as construed in classical game theory, entails common knowledge of the following claims: (a) each of the premises of the argument in Section 2 is true, (b) the premises of that argument entail that the first player is rationally obligated to choose strategy 1, and hence—by the "deductive closure" of knowledge under known entailment—(c) the first player is rationally obligated to choose strategy 1. And, of course, if common knowledge of mutual rationality entails *common knowledge* of claim (c), then it also entails claim (c) itself.

Now, it may well be that classical game theory often construes common knowledge of mutual rationality this way, at least implicitly. (Compare Rabinowicz's construal of common belief in "forward rationality," cited in Section 1 above and in note 7.) And if one deploys the locution 'common knowledge of mutual rationality' in a manner that presupposes such a construal, then certainly the following will be true: given the common-knowledge-of-mutual-rationality assumption *as thus construed*, the backwards-induction reasoning we set forth in Section 2 is not soritical.

We grant the point. We also recognize that some advocates of classical game theory not only might previously have deployed the locution 'common knowledge of mutual rationality' this way, but also might choose to continue thus to deploy it in the face of our discussion above. Technical usage of pre-theoretical terminology, purporting to explicate pre-theoretic usage of that same terminology, tends to become entrenched—even if it actually embodies presuppositions that conflict with pre-theoretic usage.

We maintain, however, that using the locution in this technical manner goes contrary to its *ordinary* meaning. Since the everyday notions of rational impermissibility, rational obligatoriness, and rational permissibility are vague in a way that matters vis-à-vis centipede games, so is the ordinary notion expressed by the locution 'common knowledge of mutual rationality'. (Compare our remarks, in footnote 7, about why one does well to eschew Rabinowicz's proposed explication of the notion 'common belief in forward rationality'.) So, although we grant that the backwards-induction reasoning set forth in Section 2 perhaps is not soritical, given common knowledge of mutual rationality *as often understood in game theory*, we contend nonetheless that this reasoning is indeed soritical given common knowledge of mutual rationality *as pretheoretically and common-sensically understood*.

Appendix 2: Vague Obligatoriness

The concept of rational obligatoriness does not exhibit "soritic spread" in the non-technical formulation we offered in Section 2 of backward-induction reasoning for the static centipede game, although the concept of rational impermissibility does exhibit this feature. And according to the common-sense view about rationality regarding centipede games—which we contend is also the correct view—no strategy dictating a specific stage at which to end the game is a rationally obligatory strategy. So in neither of these respects does the concept of rational obligatoriness "apply vaguely" to possible strategies in the game.

Nonetheless, in the following important respect, rational obligatoriness does indeed apply vaguely to potential strategies: each player is rationally obligated to choose *some* strategy that (i) dictates a specific stage at which to take two coins, and (ii) is very late in the sequence of strategies 1, 2, …, 101,

102. This form of rational obligatoriness is “collectivistic” rather than “individualistic,” because it pertains to the whole set of potential strategies without pertaining to any single member of that set—and, furthermore, pertains to the whole set without specifying any determinate boundaries on the range of potential strategies within that set that fall under the vague category “very late in the sequence of potential strategies.” Thus, the (vague) range of potential strategies any of which would satisfy a player’s collectivistic rational obligation coincides with the (vague) range of potential strategies that constitutes the “transition zone” of potential strategies each of which counts, individually, as penumbral between (i) rationally impermissible strategies that dictate taking two coins too early and (ii) rationally impermissible strategies that dictate taking two coins too late.

Although the notion of rational obligatoriness is not locally vague with respect to a single static centipede game, nonetheless it is locally vague with respect to a sequence of successive centipede games in which the first game has only one stage, the second game has two potential stages, etc. In the first game, the strategy of taking two coins at the first (and only) stage is rationally obligatory; and, as one commences through the sequence from one game to another, the category of rational obligatoriness applies locally vaguely, in a way that exhibits “soritic spread,” to the strategy of taking two coins at the first stage. Collins and Varzi (2000) focus their discussion on such sequences of successively longer centipede games, calling any game in which it is rationally obligatory to take two coins at the first stage a “take-it game.” They claim—rightly, we maintain—that the predicate “is a take-it game” is vague with respect to such a sequence of games—which reflects that fact that rational obligatoriness applies to the take-two-at-stage-1 strategy in a manner that is *locally* vague vis-à-vis the successive games in the sequence.

Appendix 3: Unsharpenability

Collins and Varzi, treating it as obvious that backwards-induction arguments concerning centipede games are soritical (and we ourselves agree), maintain that the soriticality of such arguments yields a consequence distinct from any of the apparent consequences we urged in Section 4: viz., that certain popular treatments of the logic and semantics of vagueness—notably, *supervaluationism*—cannot accommodate the kind of vagueness exhibited by notions like rational impermissibility. Supervaluationism rests on the idea that the truth value of a sentence deploying vague predicates is determined by the classical truth values that the sentence receives under the various ways in which those predicates can conceivably be “sharpened”: the original sentence is true if it is assigned True under every conceivable sharpening; is false if it is assigned False under every conceivable sharpening; and otherwise is neither true nor false.

Collins and Varzi argue that the vague notion of rational obligatoriness has no conceivable sharpenings when applied to a sequence of progressively longer centipede games. In our view, their argument for this conclusion moves too fast, because it slides between the highly plausible assumption (i) that a conceivable sharpening should be one that could be *deployed by the pertinent linguistic community at large*, and the much less plausible assumption (ii) that a conceivable sharpening should be one which, if deployed by the pertinent linguistic community at large, would mark a *publicly knowable divide* between the items in a sorites sequence that fall under the now-sharpened-concept and those that do not. What they (persuasively) argue, as we understand them, is that the notion of rational obligatoriness has no conceivable sharpenings that meet condition (ii). (The argument is by backwards induction.)

Condition (ii), however, seems too strong to be plausible. A suitably plausible constraint on the notion of sharpenability, we maintain, is (i) rather than (ii). A *highly* salient way to meet condition (i), without also meeting condition (ii), would be for all members of the linguistic community to agree to a single population-wide sharpening that yields *agent-relative* sharp boundaries for the concepts of rational obligatoriness, rational permissibility, and rational impermissibility—where each agent’s boundaries are determined by some specific subjective-probability distribution (known to that agent) and some specific subjective-utility assignment (also known to that agent). Such a sharpening could be deployed, knowingly, by the linguistic community at large—even though each member of the community need only know her or his own (sharpening-determined) subjective probabilities, subjective utilities, and expected-utility-maximizing options—not those of other people.

As regards the static version of the centipede game described in the first paragraph of the present paper, both players could know that each of them has some specific probability distribution, over the other’s available strategies, that dictates (by expected-utility maximization) a specific strategy for oneself—without either player knowing what the other’s probability distribution is or what strategy available to the other has maximal expected utility for the other. (And both players could be in this epistemic situation even with the usual common-knowledge assumptions in force.)

So we find the Collins-Varzi argument unpersuasive. But is there a way to modify their dialectical strategy and thereby generate a sound argument in support of their claim that rationality notions cannot be sharpened vis-à-vis centipede games? Perhaps so, although the reasoning requires certain supplementary—yet plausible—assumptions about rational agents who have common knowledge of both mutual rationality and mutual self-profit-maximizing motivation. Such an argument, for the static centipede game discussed above, might go as follows:

> Assume, for reductio, that for each player there are multiple permissible potential probability distributions over the other player's available strategies—where each permissible distribution assigns non-zero probabilities only to strategies, other than strategies 99-102, that are *late* on the strategy list in section 1.5 above. Assume too that there is common knowledge that each player will adopt a specific probability distribution over the possible strategies available to the other player and will select whatever strategy maximizes (under that probability distribution) her/his expected utility. Given the usual common-knowledge assumptions (viz., common knowledge of mutual rationality, and common knowledge that each player is motivated solely by the goal of maximizing her/his profits in the current centipede game), if a player P tentatively adopts a particular probability distribution D over propositions concerning which available strategy the other player would follow in the game, then when P considers all the various possible probability distributions (concerning this same matter) that the other player might tentatively adopt concerning P herself/himself, P's *expected value* for the other player's tentative probability distribution should be a probability distribution D# that matches D. (Matching means that the probability that D# assigns to P's i-th available strategy is identical to the probability that D assigns to the other player's i-th available strategy, for each of the successive 51 strategies available to a given player.) That should lead P to replace D by a new tentative probability distribution D*—where the strategy for P that maximizes P's expected utility under D* is more conservative than the strategy that maximizes P's expected utility under D. This reasoning iterates repeatedly, ultimately yielding the conclusion that P must assign probability 1 to the strategy of taking two coins at the very beginning—which contradicts the assumption.

As formulated, this argument assumes for simplicity that the successive tentative probability distributions would always maximize the expected utility of some *single* strategy. This assumption can be relaxed, however; now the key idea is that each successive tentative probability distribution D* should yield a *set* of expected-utility-maximizing strategies each of which is more conservative than any corresponding member(s) of the set of such strategies that was yielded by D.

We ourselves find this modified version of the Collins-Varzi argument quite plausible. So yet another apparent consequence of the soriticality of backwards-induction reasoning in centipede games is Collins and Varzi's contention that the vagueness of practical-rationality notions is unsharpenable with respect to such games.[13]

References

Aumann, R. (1995). "Backward Induction and Common Knowledge of Rationality," *Games and Economic Behavior* 8: 6-19.

Binmore, K. (1987). "Modelling Rational Players: Part 1," *Economics and Philosophy* 3: 179-213.

Bicchieri, C. (1988). "Common Knowledge and Backward Induction: A Solution to the Backward Induction Paradox. In M. Vardi (Ed.), *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*. Los Altos: Morgan Kaufman Publishers: 381-393.

Bicchieri, C. (1989a). "Backward Induction without Common Knowledge," *Proceedings of the American Philosophical Association* 2: 239-243.

Bicchieri, C. (1989b). "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," *Erkenntnis* 30: 69-85.

Bicchieri, C. (1993). "Counterfactuals, Belief Changes, and Equilibium Refinements," *Philosophical Topics* 21.1: 21-52.

Binmore, K. (1994). "Rationality in the Centipede," in R. Fagin (ed.), *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Fifth Conference (TARK 1994)*. San Francisco: Kaufmann, 150-159.

Broome, J. and Rabinowicz, W. (1999). "Backwards Induction in the Centipede Game," *Analysis* 59: 237-242.

Collins, M. and Varzi, A. (2000). "Unsharpenable Vagueness," *Philosophical Topics* 28: 1-10.

Horgan, T. (1994). "Robust Vagueness and he Forced-March Sorites Paradox," *Philosophical Perspectives* 8: 159-188.

Horgan, T. (2010). "Transvaluationism about Vagueness: A Progress Report," *Southern Journal of Philosophy* 48: 67-94.

Horgan, T. (2017). "Troubles for Bayesian Formal Epistemology," *Res Philosophica* 94: 233-255.

McKelvey, R. and Palfrey, T. (1992). "An Experimental Study of the Centipede Game," *Econometrica* 60: 803-836.

Nagel, R. and Tang, F. (1998). "Experimental Results on the Centipede Game in Normal Form: An Investigation on Learning," *Journal of Mathematical Psychology* 42: 356-384.

Pettit, P. and Sugden, R. (1989). "The Backward Induction Paradox," *Journal of Philosophy* 86: 169-182.

Priest, G. (2000). "The Logic of Backwards Inductions," *Economics and Philosophy* 16: 267-285.

Rabinowicz, W. (1998). "Grappling with the Centipede. Defence of Backward Induction for BI-Terminating Games," *Economics and Philosophy* 14: 95-126.

Reny, P. (1989). "Common Knowledge and Games with Perfect Information," *Proceedings of the Philosophy of Science Association* 2: 363-369.

Rosenthal, R. (1981). "Games of Perfect Information, Predatory Pricing, and the Chain Store," *Journal of Economic Theory* 25: 92-100.

Sobel, H. (1993). "Backward Induction Arguments in Finitely Iterated Prisoners' Dilemmas: A Paradox Regained," *Philosophy of Science* 60: 114-133.

Sorensen, R. (1988). *Blindspots*. Oxford and New York: Oxford University Press.

Stalnaker, R. (1996). "Knowledge, Belief and Counterfactual Reasoning in Games," *Economics and Philosophy* 12: 133-163.

Stalnaker, R. (1998). "Belief Revision in Games: Forward and Backward Induction," *Mathematical Social Sciences* 36: 31-56.

Stalnaker, R. (1999). "Extensive and Strategic Forms: Games and Models for Games," *Research in Economics* 53: 293-319.

[1] See, for instance, Rosenthal (1981), Binmore (1987, 1994), Aumann (1995). Pertinent experimental work includes McKelvey and Palfrey (1992), Nagel and Tang (1998). Philosophical discussions include Bicchieri (1988, 1989a,

1999b, 1993), Sobel (1993), Sorensen (1988), Pettit and Sugden (1989), Stalnaker (1996, 1998, 1999), Rabinowicz (1998), Reny (1989), Broome and Rabinowicz (1999), Collins and Varzi (2000), Priest (2000), Smead (2008), Baltag, Smets, and Zvesper (2009).

[2] For different formulations of this objection see, for instance, Binmore (1998), Reny (1999), Bicchieri (1989a), Pettit and Sugden (1989.

[3] One way to try reconstructing the first player's rationale for ending the game on the first move would be to construe the reasoning as involving two sequential segments, as follows. In the first segment one says to oneself,

> I know that both of us will retain our common belief in mutual rationality as long as neither of us has yet made an irrational move; this entails that there is some stage n such that the game will end at stage n without either player having yet made an irrational move; and this in turn entails that the pertinent material-conditional statements that figure in the Rabinowicz proof are all true. (They are all true because (i) the corresponding counterfactual conditionals, expressing resiliency and robustness up through stage n, are all true up through stage n, and (ii) the material-conditional statements whose antecedents pertain to stages later than n are all vacuously true by virtue of having false antecedents.)

One also says to oneself, "My reasoning thus far does not tell me, of any *specific* stage n, that the game will stop at stage n without either player having made an irrational move; rather, it only tells me that *there is* some such stage." In the second segment, one invokes backwards-induction reasoning that appeals, successively backwards from stage n, not to the pertinent material conditionals but rather to the corresponding counterfactual conditionals concerning what the players *would* do were the game at stage n, were it at stage n-1, and so on backwards—thus leading to the conclusion that the first player not only *will* end the game on the first move, but is rationally required to do so.

A serious problem with this approach, however, is its presumption that one can legitimately invoke backwards induction without knowing which stage in the backwards-inductive sequence would be the *first* stage. That presumption looks very dubious, and indeed goes contrary to standard thinking in game theory about acceptable backwards induction.

[4] Another way to try reconstructing the first player's rationale for ending the game on the first move, different from the approach described in the preceding footnote, is suggested by the following remarks from Rabinowicz. He is discussing a centipede-like game called Take It or Leave It, where there is a first player X and a second player Y:

> Suppose X believes that, if he went across, the game would continue for a short while, with both players moving across, and then stop with Y taking the pot at, say, the fourth or the sixth node. Then, in view of our proof above, he *cannot* expect the conditions of forward rationality to obtain at the intermediate choice nodes (such as numbers two and three)… In particular, the first player might expect that the player of the third node (X himself) either would not act rationally (violation of resiliency) or would not be confident of

> his opponent's rationality at the subsequent node (violation of robustness). It follows, then, that the backward-induction behavior under conditions of forward rationality can be rationalized without ascribing to him the belief that these conditions would invariably survive counterfactual developments. (pp. 112-113)

The final sentence in this passage expresses the specific point that Rabinowicz is making, which is correct. But, with this passage in mind, one might consider embracing the following general claim about what rationality requires: in centipede games and in structurally similar games like Take It or Leave It, the first player's rationale for stopping the game on first move consists in the fact that he believes, of some specific stage n, that (a) the player at stage n *would* stop at stage n if the game were to reach that stage and (b) robustness and resiliency hold with respect to stages 1 through n.

One worry about this proposal is that a belief of type (a) is evidentially unwarranted relative to the body of evidence that is stipulated to be available to the players. Another worry is this: the lower n is, the closer is one's belief of type (a) to begging the question at issue about rational play (since the common-sense view is that it is rationally impermissible to end a centipede game before very near the final potential stage); whereas the higher n is, the stronger and more questionable are the needed resiliency assumptions of type (b).

[5] The policies we are calling 'strategies' sometimes are called 'strategy profiles', especially when dynamic games are being considered. In that alternative terminology, each individual move at any given stage, within a total policy for playing the whole game, counts as a strategy. But our coarse-grained use of 'strategy' seems more apt for the static centipede game now under consideration.

[6] The same goes for the concept of rational obligatoriness, which is surely sorites-susceptible too. However, the structure of static centipede games prevents this concept from "soritical spread" *within such games*: if any single strategy is rationally obligatory in a centipede game, then all others are rationally impermissible. Rational impermissibility is the rationality-concept that exhibits soritical spread in these games. (In Appendix 2 and Appendix 3 we say more about vagueness and rational obligatoriness vis-à-vis these games.)

[7] For the same reason, when one considers a sequence of centipede games in which the first game has only two possible stages and each successive game has one more possible stage than its predecessor, there is no definite fact of the matter about which such games are ones for which a rationally optimal strategy dictates cooperation rather than defection for the first stage.

[8] Yet another theoretical advantage of the vagueness hypothesis deserves mention. As we noted in Section 1, Rabinowicz (1998) shows that the conclusion that the first player *will* end the game on the first move is derivable, by backwards induction, from the assumption that the first player will make a rational move plus the plausible-looking assumption that the first player has a material-conditional "forward rationality" belief that iteratively embeds a succession of claims about other material-conditional beliefs: the first player believes that his successor (if

there is to be one) both (1) will make a rational move, (2) believes that *her* successor (if there is to be one) both (2.1) will make a rational move and (2.2) believes that *his* successor (if there is to be one) both (2.2.1) will make a rational move and (2.2.2)…, etc. (Many—or most, or all—of the progressively embedded material conditionals might be vacuously true, by virtue of having false antecedents.) Yet, as also noted in Section 1, Rabinowicz goes on to observe—correctly—that this argument does not constitute a *rationale* for stopping the game on the first move, since such a rationale would require the first player to hold appropriate counterfactual beliefs about what *would* happen in the continuation of the game if he were to act otherwise than stopping it on the first move. Now, something seems wrong here. How could assumptions about *rationality*—specifically, about common knowledge of forward rationality—validly generate the conclusion that the first player will end the game on the first move without thereby also providing a rationale for that very conclusion?

If the vagueness hypothesis is correct, then this doesn't happen after all. Although it is true, given the background assumption of common knowledge of mutual rationality, that the players both believe that they will retain their common belief in mutual rationality as long as neither player has yet made an irrational move, this belief is not identical to—and does not entail—the iteratively complex, material-conditional, belief that Rabinowicz attributes to the first player. Since the category *making a rational move* is vague, so is the category *believing that one's successor (if there is to be one) will make a rational move*. Thus, to hold that the first player's game-initial belief in mutual forward-directed rationality consists in the first player's initially holding the Rabinowicz-attributed belief would be tantamount to embracing a sorites argument regarding the content of that initial forward-rationality belief—which, if indeed the category *rational move* is vague, would be a mistake.

[9] How, specifically, does one reject a principle that functions as a culprit-assumption in a sorites argument? There are various candidate answers to this question, reflecting various competing proposed treatments of the logic and semantics of vagueness. Using the category *heap* for illustration, we here sketch two potential answers. As a prelude, we reformulate schema $\mathcal{H}$ of Section 3 as a conjunction of conditionals rather than a schema. (One could do the same with the schemas $\mathcal{C}_{A,B}(\alpha)$ and $\mathcal{C}_{B,A}(\beta)$.) Letting 'H(i)' symbolize 'A pile of sand containing n grains is a heap', and letting the variable 'n' range over natural numbers i such that $10^7 \geq i \geq 20$:

H: $[H(10^7) \supset H(10^7\text{-}1)] \;\&\; [H(10^7\text{-}1) \supset H(10^7\text{-}2)] \;\&\; \ldots \;\&\; [H(21) \supset H(20)]$

According to supervaluationism (the most popular approach among philosophers to the logic and semantics of vagueness), the repudiation of **H** should go as follows. First, one affirms the classical negation of **H**:

~H: $\sim\{[H(10^7) \supset H(10^7\text{-}1)] \;\&\; [H(10^7\text{-}1) \supset H(10^7\text{-}2)] \;\&\; \ldots \;\&\; [H(21) \supset H(20)]\}$

This blocks a sorites argument that uses **H**. Second, one affirms the following statement, which (according to classical logic, and also under supervaluationist semantics) is logically equivalent to statement **~H**:

E: $\{[H(10^7) \;\&\; {\sim}H(10^7\text{-}1)] \vee [H(10^7\text{-}1) \;\&\; \sim H(10^7\text{-}2)] \vee \ldots \vee [H(21) \;\&\; {\sim}H(20)]\}$

(We sketch supervaluationist semantics in Appendix 3.) Third, one denies that any disjunct in **E** is true. Fourth, one claims that the fact that **E** contains no true disjunct is enough to honor the vagueness of the notion of rational impermissibility.

An alternative approach (Horgan 1994, 2010), which has the advantage of honoring the fact that statement **E** *seems* to affirm the existence of a sharp boundary between being rationally impermissible and being not rationally impermissible, goes as follows. First, claim that **H** is neither true nor false, and likewise for **E**, and likewise for their classical negations **~H** and **~E**. Second, introduce a non-classical negation-operator, '⌐', which works semantically this way: ⌐Φ is true iff Φ is not true—i.e., iff either Φ is false or Φ is neither true nor false. Third, affirm the non-classical negations of **H**, **E**, **~H**, and **~E**. By affirming **⌐H** one blocks a sorites argument that uses **H**; by affirming **⌐E** one avoids the counterintuitiveness of embracing **E**; and by non-classically negating each of the statements **H**, **E**, **~H**, and **~E**, one "logically quarantines" all four of these statements, thereby preventing any of them from becoming available to feed logically valid inferences that lead to paradoxical conclusions.

[10] A referee suggests that an adherent of expected-utility maximization might say, rather, that the player has *vague* probability assignments over the available strategies of the other player, and thus that it becomes vague which of these probability assignments maximizes the given player's expected utility—which leaves intact the definition of rational obligatoriness/permissibility as expected-utility maximization. This dialectical move remains open, and we have not precluded it here. The move does encounter the following apparently significant worry, however: saying that it's vague which probability assignment maximizes "the" expected utility of a given strategy for the given player is apparently analogous to saying that it's vague what constitutes "the" sharp transition between heaphood and non-heaphood. Just as the definite description

the sharp transition between heaphood and non-heaphood

has no referent (the worry goes), likewise a definite description of the form

the expected utility of strategy S for player P

apparently has no referent either. This threatens to undermine the very intelligibility of the idea that it's a vague matter what constitutes "the" expected utility of strategy S for player P.

[11] We ourselves are inclined to embrace pessimism. For argumentation in support of pessimism, see Horgan (2017).

[12] This objection was raised by a referee.

[13] For helpful comments on ancestors of this paper we thank Aaron Bronfman, Juan Comesana, Brian Fiala, an anonymous referee, and audiences at the University of Alabama, the University of Auckland, the University of Delaware, and the University of Nebraska.