Jim Joyce, "The Role of Incredible Beliefs in Strategic Thinking" (1999)

Prudential rationally is a matter of using what one believes about the world to choose actions that will serve as efficient instrument for satisfying one's desires. Much of my research concerns the role that beliefs play in this process of rational decision making. This currently an active area of investigation has engaged the effort psychologists, computer scientists and economists, as well as philosophers. In this essay I want to consider a special category of beliefs, beliefs about the incredible, that concern what would or will happen if events that a decision maker is absolutely certain will not occur do end up occurring. Philosophers have recognized the importance of subjunctive beliefs in rational decision making for some time, and it is now widely a decision maker's beliefs about subjunctive conditionals of the form, "If I were to perform such-and-such act then such-and-such an outcome would follow," are crucial to understanding what should be done in a given situation. The sort of "incredible" beliefs I am going to discuss here are a species of these subjunctive beliefs, but with two twists. First, they concern events that the decision maker is absolutely certain will not transpire, which is not true of all counterfactual beliefs. Second, the decision maker is required to treat these events not merely as possibilities to be supposed true in a purely hypothetical way, but as potential items of information that could be learned. My aim here is to explain why a complete account of rational decision making requires an analysis of such "incredible" beliefs. I shall argue, in particular, that decisions involving strategic interaction, of the sort that game theorists find interesting, cannot be understood unless we suppose that decision makers have beliefs about the incredible. To illustrate the point I will consider a famous decision problem known as the Centipede game, which some game theorists see as generating insoluble paradoxes. In fact, these paradoxes dissolve as soon as we introduce "incredible" beliefs into the picture.

Game theory deals with decision problems of strategic interaction in which the question of what a person ought to do depends on his ability to emulate another person's reasoning and thereby to anticipate her choices. Consider a situation that arises in our Department every time we have a visit by an outside speaker. Let's call it the Hosting Game. It is our custom to take speakers out to dinner after their talks. One faculty member is designated as the official "host," a mildly annoying job since the host takes care of the restaurant bill and must wait a month or so to be reimbursed by the University. About a week prior to the speaker's arrival, our Chairperson sends out an e-mail to the faculty asking who wants to attend, and whether anyone is willing to host. Answering such messages involves strategic thinking of an exceedingly high order of complexity. First, there is the matter of what response send. One can volunteer to host and risk getting stuck with the bill, but there is also a chance that your offer will arrive after someone else has volunteered in which case you build up credit with the Chair for being a "team player" without actually having to host. Another alternative is to ask to attend without volunteering to host. The problem here is that if no one volunteers then you risk of getting stuck with the job. Finally, one can claim to be busy, but then one cannot attend dinner. Timing presents further complications. Seating at dinner is limited, which mitigates in favor of getting one's request in early. The rub is that if no host comes forward the first respondent is likely to get stuck with a job. The best possible outcome would be send a response early enough to be included on the dinner list, to volunteer to host, thereby building up credit with the chair, but to receive the reply, "Thanks for your kind offer, but Professor X has already volunteered." You can imagine how much time we all have spend on this!

Of necessity, our reasoning is strategic: we each choose a course of action based on our best guesses about the actions of our colleagues. We make these guesses by trying to reason things out the way the others would, given what we know about them and what we think they know about us. There are many wrinkles to consider: Does Velleman expect Anderson to reply before 2:00pm if she believes that Gibbard or Proops will have replied by then? If so, will Velleman volunteer at 2:10pm if he thinks Curley believes that Hills will volunteer at 2:08pm? Does Sklar suspect Railton of suspecting Ivanhoe of suspecting me of offering to host if I think Tappenden is going to offer? Will Lormand and Loeb decide to reply if they believe that Thomason is not volunteering? Might Walton want to host? Might Darwall decide to host the dinner himself if no one replies by 5:00pm? Does Hofweber yet know that he picks up the check if he hosts, and will his ignorance of the "rules of the game" lead him to offer? With all these weighty matters to consider it’s a miracle that we get any teaching and research done at all!

It should be clear that the right course of action for any of us depends on how much we know about the others. The more information we have about our colleagues motives, habits and beliefs (especially their beliefs about the motives, habits and beliefs of other colleagues), the better positioned we are to make a wise decision. In real life, questions about the extent of a person’s knowledge about the beliefs and motives of others are involved and messy. Rather than get into these complications, game theorists adopt the idealizing assumption that they are dealing with individuals who have a great deal of knowledge about those with whom they interact. Specifically, game theorists usually assume that they are dealing with a decision problem, a "game" as it is called, in which: (a) all players are rational; (b) all players understand the structure of the game; (c) all players know what the others believe and want; (d) all players know (a)-(c); (e) all players know (d); (f) all players know (e); ex cetera. Let’s refer to these the common knowledge (CK) assumptions. Nearly all of classical game theory is premised on the idea that the CK assumptions are satisfied.

The resulting theory is exceedingly rich in consequences. One key result is the Equilibrium Theorem, which states that in any game in which the CK assumptions are satisfied the players’ choices will instantiate a Nash equilibrium in which no player would have an incentive to change his action even if he were to discover what all the others were going to do. A Nash equilibrium is often described as a "self-enforcing agreement" in which each player’s act is a best response to the acts of other players. Some games have more than one equilibrium, and the question of precisely which one a group of rational players will settle on has been intensively investigated over the past thirty years. I am going to leave these complications aside, however, and focus on games with a single equilibrium. For games of this sort the game-theorist’s criterion of rationality is unequivocal: players who satisfy the CK assumptions will play their end of its unique equilibrium. This requirement can be given the following rationale: Since each player is rational and knows everything there is to know about the others’ beliefs and motivations, each should be able to predict what the others will do by putting herself in their shoes and emulating their reasoning. Thus, each player should be able to deduce in advance what the others will do and be able to base her decision on this information. But, if all the players know in advance what the others are going to do then, being rational, they will all choose actions that are best replies to what the others do. Thus, players who satisfy the CK assumptions will always make choices that instantiate a Nash Equilibrium.

Using this basic idea and variations on it, game theorists have been able to "solve" a great many decision problems. One might wonder, however, whether there is any point to this exercise given that the CK assumptions on which the whole edifice is based are idealizations that are far removed from what one finds in real life. What's the point of constructing an elaborate theory of decision making for ideal agents who are nothing at all like human beings? As many examples from the history of science attest, there are two good reasons for developing theories that apply to ideal situations before trying to treat more realistic cases. First, it is sometimes easier to the handle realistic cases when one has a theory for ideal ones in hand. Second, it often possible and useful to view non-ideal cases as approximations of ideal ones. Things can work out this way in game theory. Once we know how perfectly rational agents will behave under conditions of complete knowledge we usually have an easier time saying how less perfect agents should act under conditions of less complete information. Likewise, in many instances the right thing for non-ideal agents to do closely approximates what ideal agents do in similar circumstances, and the smaller the deviation from the CK assumptions the better the approximations of the ideal theory tends to be. Still, this is not universal; some of the most fascinating games are those in which small deviations from the common knowledge assumptions give rise to great disparities in what players should do. This is beautifully illustrated by the "Centipede" game.

Suppose that Hera and Zeus, two ideal agents, are seated at a table on which there is a small cup of nectar and two piles of envelopes marked $1,000, $2,000, $3,000, and so on to $100,000. Each envelope holds the amount marked on it in cash. The rules of the game are simple: Each player must pay a $1 entry fee to play. Hera, who gets the first move, can either take her $1,000 envelope and forgo the nectar, in which case it becomes Zeus' turn, or she can take the envelope and drink the nectar, in which case the game ends. Zeus has the same options if he gets a turn. The two gods go back-and-forth in this way, each taking the least valuable of their remaining envelopes, until one drinks the nectar, or until the envelopes run out. Let's assume that our deities satisfy the following conditions:

Both like money to about the same degree that an ordinary middle-class American does. (So, $1,000 is a desirable prize, $100,000 is a wonderful one.)
Each is interested only in maximizing his or her own fortune: Zeus does not care how much money Hera makes, and Hera does not care how much Zeus makes.
At every point in the game both Zeus and Hera will be mildly thirsty and will prefer drinking the nectar to not drinking it. Still, they always prefer having an extra dollar to having the nectar.
Hera and Zeus satisfy the common knowledge assumptions: They both know the facts just stated, the rules of the game, and both are convinced that the other is rational. Moreover, they both know they both know all this, and they both know they both know that both they know all this, and so on.
Just before the game begins, Zeus offers you the chance to share his winnings, fifty-fifty, if you will just stake him his entry fee. You know that Zeus is perfectly honest, perfectly rational, and that he and Hera satisfy the CK assumptions. Should you give Zeus the dollar he needs to play the game?

When first hearing about Centipede, nearly everyone is willing to put up the fee, which seems a pittance when compared to the possible winnings. The only smart play for Zeus and Hera, it seems, is to forgo the nectar and to do a bit of self-interested "sharing" until near the end, when one or the other will bail out and drink. It appears to be in both their interests to do this. After all, if they can only live with a mild thirst for forty-seven rounds they will be millionaires. If they can get to round ninety-five they will each be worth more than 4.5 million. And, you get half Zeus’s take! Why not enter and get rich?

Unfortunately, it's a sucker bet; you will end up losing your dollar. Players like Zeus and Hera can never get past the first round of Centipede; the game’s only Nash equilibrium is the one in which both players drink the nectar the first chance they get. There is an airtight argument, the backward induction argument, which shows that this is so. Its logic is relentless and inescapable. Consider Hera’s final turn. This will be her last chance to drink, so she will be faced with a straight, non-strategic choice between taking $100,0000 and forgoing nectar or taking the $100,000 and drinking it. Since she does not care about how much Zeus wins, she will surely drink. Since both players satisfy the CK assumptions Zeus will be able to figure this out just as easily as we did, so he will surely drink if the game gets as far as his next to last turn. Hera can deduce this, so she will surely drink if the game gets as far as her next to last turn. But, Zeus can deduce this as well, so he will drink on his third to last turn. I am sure we can all see which way the wind is blowing here; backwards induction leads inexorably to the conclusion that Hera will stop the game on her first turn, leaving Zeus and you with zilch, and $10,000,000 on the table.

To make matters worse, this reasoning holds up even if there is ten billion dollars at stake rather than ten million. As long as Zeus and Hera have common knowledge of one another's beliefs, motives, rationality, and the structure of the game, both can reason to the conclusion that she should stop the game on her first turn. In fact, even if the deities played Centipede a million times in succession backward induction would still dictate that Hera should bail out in the first round of singly every game. (It is an instructive exercise to work out why.) Not even pregame communication can save them. Before the game starts Zeus will happily agree not to touch the nectar until the last round of play if Hera doesn't. But, she will take this for what it is: an incredible promise that she cannot force Zeus to keep. Surprisingly, things are no better in a finitely repeated version of Centipede even though it might seem that there Hera can threaten Zeus with retribution for failing to keep his promise. "If I let you have a turn in the first game and you don't keep the game going," she will sternly announce, "I will punish you by ending the second game on my first turn." Unfortunately, this is not going to move Zeus. Since he appreciates the force of backwards induction reasoning, he will recognize Hera’s threat as inert not because she won't carry it out, but because she will carry it out whatever he does. (Again, it’s instructive to work out why.) In the end, the only conclusion to draw is that Hera and Zeus, rational agents who know everything relevant about the decision situation they are in, can never get rich playing Centipede.

Nearly everyone is incredulous when they first hear this. Can it really be that rational players will let more than ten million dollars go to waste? Given that it is in both their interests to continue the game, and since the "prize" for ending it is so trifling, it is hard to believe that they will not be able to think their way past the first round. Even if backward induction has a kind of "formal" correctness, it seems absolutely crazy to follow its dictates when real money is on the table!

There is more than a grain of truth in this. Centipede is one of those games in which the right acts for ideal agents differ radically from those for even slightly less than ideal agents. If you and I played the game we would not end up poor like Zeus and Hera because we would not satisfy the CK assumptions, which are crucial to the success of the backward induction argument. Its highly counter-intuitive conclusion cannot be drawn if Hera is not convinced that Zeus can follow backward induction reasoning, or if Zeus is not convinced that Hera is convinced of this, or if one of them is not sure the other wants the nectar, or if Zeus suspects that Hera might be concerned about his welfare, or if Hera suspects that Zeus suspects this. Economists have examined these matters at length and, without going into the details, the upshot is that nearly every weakening of the common knowledge assumptions, even modest ones, undermines backward induction in Centipede. Rational players with less than complete knowledge of one another's beliefs, motives or reasoning powers end up cooperating deep into Centipede and getting rich.

It is easy to see how even a little uncertainty can have a powerful effect here. Suppose that Hera suspects that Zeus suspects that she might be interested enough in his welfare to forgo the nectar on any given round. It would then be in Hera’s interest to confirm Zeus’s suspicions so as to give him a reason to extend the game if he gets the chance. Fortunately, she has the perfect instrument at her disposal: by not drinking on her first turn she will reinforce Zeus’s belief that she has his interests at heart. Thus reinforced, Zeus will be inclined to let Hera have a second turn. This will confirm her suspicion that he believes that she has his interests at heart, which gives her further reason to keep the game going. As the game continues a "feedback loop" is established in which the act of forgoing the nectar by one player always reinforces the beliefs that make it rational for her opponent to forgo the nectar in the next round. Of course, for such a process to work the game must be sufficiently long (e.g., it will not succeed if there are only four envelopes on the table), but as the number of envelopes gets larger and larger the amount of uncertainty it takes to produce an extended game shrinks rapidly. For games with a hundred rounds, like Centipede, it takes only a minuscule deviation from CK to undermine backward induction.

This should go a long way toward assuaging our concerns about Centipede. Once we appreciate how much common knowledge backward induction requires, and how unstable it becomes under small changes in this knowledge, it ceases to be so troubling that players who satisfy the CK requirements end the game before it really starts. This strikes us as the wrong result only because we, and those with whom we interact, are far removed from the ideal that backward induction presupposes. Once we understand that the reasoning requires players who know everything there is to know about one another and the game the backward induction solution seems less mad. Centipede provides one of those cases in which the theory of rationality for ideal agents is not a reliable guide to what less than ideal agents, like all of us, should do. The fact that Zeus and Hera will not get rich if they behave rationally tells us nothing about what will happen to us if we act rationally.

Still, we should not think that the paradox has been resolved. There remains something deeply troubling about Centipede. First, when the game is truncated, so that the piles contains six envelopes each, even agents like you and I should behave just like ideal agents (so long as we think the other player is at least moderately rational and only out to maximize his own profits). Moreover, even in the hundred-envelope version it is still perplexing that you and I, with our limited knowledge and rationality, can get rich playing the game while Zeus and Hera, perfectly rational beings with perfect knowledge, cannot? This seems particularly odd when one reflects on the fact that both Zeus and Hera will recognize that only their knowledge stands in their way. "If only we did not know so much," they might lament, "we could secure our happiness as lesser beings do. But alas, since we do satisfy the CK assumptions we both know that the only rational play for Hera is to end the game on her first turn." The strange thing is that this lament seems to contain its own solution. Since both players know that the CK assumptions can only be satisfied if Hera ends the game on her first turn, it follows that if Hera does not do this then both players can be sure that the assumptions are not satisfied. Thus, since Hera is free to do as she pleases, it appears that by not drinking the nectar she has the power to provide Zeus with evidence that will conclusively undermine his belief that the CK assumptions hold. Refraining from drink is what game-theorists call a counter-theoretical, an act that no rational player with the amount of common knowledge game theorists assume would ever commit. What makes Centipede so interesting is that it seems like the counter-theoretical act is actually the rational act! After all, by not drinking Hera appears to position herself to make millions while drinking forces her to settle for $1,000. This is the paradox of backward induction: if the CK assumptions hold, then Hera is sure to act rationally, but it seems that she can only act rationally by ensuring that the CK assumptions do not hold.

Philosophers and economists have had a lot to say about this paradox. Some take it to show that the game theoretic conception of rationality undermines itself. Others claim that agents who satisfy the CK requirements are not really free to perform counter-theoreticals; their great knowledge is supposed to somehow deprive them of even the capacity to act irrationally. Both suggestions are misguided. We will soon see that, contrary to what it seems, if the CK assumptions hold then Hera will know that she cannot make herself better off by forgoing the nectar. A third response to the paradox is more promising. I call it the "theorists" response. Game theorists have traditionally seen their task as one of prediction, of finding methods that reliably predict the behavior of players who satisfy the CK assumptions. If this is the goal then the counter-theoretical question of what would happen if Hera refrained from drinking can be ignored, and the backward induction "paradox" vanishes. According to the theorist, what the backward induction argument shows is that Hera not drinking is logically inconsistent with the CK assumptions. This makes asking what would transpire if the CK assumption hold and Hera does not drink akin to asking what the radius of a circle would be if it were also a square. The theorist can legitimately refuse to answer on the grounds that it is impossible to evaluate a counterfactual conditional with a contradictory antecedent.

There is something right in this response, but something lacking as well. When thinking about a person’s behavior from the theorist’s perspective one asks about the antecedent conditions that causally explain the action and allow us to predict its occurrence. This is a worthwhile endeavor, but we must keep in mind that there is a significant difference between predicting acts and justifying them. The theorist’s perspective is to be contrasted with the agent’s perspective, where the aim is not prediction and causal explanation, but justification and rationalizing explanation. Here one takes the agent’s point of view and asks what it is about her beliefs and desires that makes the act reasonable for her. To fully resolve the paradox of backward induction we need to see Hera’s act of drinking the nectar as one that makes sense from her perspective. This is something the theorist’s response cannot supply. Merely being told that drinking is inconsistent with the CK assumptions leaves us completely in the dark about Hera's reasons for doing as she does.

Now, it might be objected that Hera and Zeus could justify their acts using the same argument that the theorist uses to predict them. Can't they simply reason that since they are rational and satisfy the CK assumptions, and since drinking is the only rational strategy for Centipede players who satisfy the CK assumptions, it follows that they should drink? If Hera and Zeus did reason this way they would be committing what I call the fallacy of presumption. This occurs whenever a person appeals to her own rationality in the course of justifying one of their actions. The problem with such appeals is that they render the justification circular. In this context a justification can be thought of as an argument whose premises exhibit the agent's reasons for doing what she does and whose conclusion says that she should do it. The agent counts as acting rationally just in case her reasons are good ones. But, if one of the premises states that the agent acts rationally, then her justification begs the question since the only way for her to (non-circularly) establish the claim that she acts rationally is by showing that she acts for good reasons. Thus, while there is no problem with Hera appealing to both her own and Zeus's rationality in predicting that she will drink the nectar, and no problem with Hera appealing to Zeus's rationality when trying to justify drinking it, it would be fallacious for Hera to appeal to her own rationality in this justification. To justify her act Hera must show how drinking serves her interests given what she wants and believes. The fact that she is rational, even if she is certain of it, cannot figure essentially in her reasons for acting.

Though this limits the players' justificatory options in one way, it expands them in another. Since neither Hera nor Zeus can assume that they themselves are rational (in the context of justification) neither can assume that the CK assumptions hold, and this lets them make sense of counter-theoretical acts. This is fortunate because Centipede players who meet the CK conditions need to be able to think about one another's beliefs and intentions under the supposition that counter-theoreticals are preformed, like the act of Hera refraining from drinking the nectar on her first turn. This is true even though both players are able to predict, with complete certainty, that these acts will not be performed, i.e., if the acts in question are "incredible" events from both player's point of view. The project of rationalizing acts is an essentially subjunctive affair. To rationalize an action is to show that all alternatives to it could be expected to lead to less desirable outcomes if they were performed. Thus, drinking the nectar is rational for Hera only if she believes that she would be better off were she to drink than were she to refrain. But, since her payoff for refraining depends on what Zeus would do in that event, and thus on what he would believe and want in that event, it follows that Hera's reasons for drinking depend on her beliefs about what Zeus would believe if he were to learn that she performed the "incredible" act of not drinking.

To clarify things, and save on ink, let H1 be the hypothesis that Hera refrains from drink on her first turn, Z1 be the hypothesis that Zeus gets a first turn and refrains from drink, H2 be the hypothesis that Hera gets a second turn and refrains from drink, Z2 be the hypothesis that Zeus gets a second turn and refrains from drink, and so on. These are all counter-theoreticals, and both Hera and Zeus will be quite certain that they are false. Yet, it is their beliefs about what would happen if these hypotheses were true, and the other player learned as much, that provide their reasons for acting. Given her desires, Hera has a sound rationale for refraining from drink on her first turn just in case she assigns non-negligible credence to the subjunctive conditional "If H1 were true, then Z1 would also be true." But, since Hera knows Zeus is rational she will assign this conditional non-negligible credence only if she also thinks that Zeus, if he were to learn H1, would assign non-negligible credence to the conditional "If Z1 were true, then H2 would also be true." Zeus will only do this if he believes that Hera, were she to learn Z1, would assign non-negligible credence to "If H2 were true, then Z2 would also be true." Continuing on in this way we may conclude that Hera is only justified in refraining from drink if, for each stage j = 1, 2,..., 100, she assigns non-negligible credence to:

If Zeus were to learn Hj then he would assign non-negligible credence to the hypothesis that Hera would decide forgo the nectar on turn j + 1 turn if she were to learn Zj.
This makes it clear that Hera has a rationale for forgoing the nectar only insofar as she believes that doing so would cause Zeus to alter his beliefs about how she will act on future turns. Since she is certain she will drink the nectar, and since she is sure that Zeus is certain of this as well, she must both suppose the truth of a proposition that she regards as certainly false and make judgements about the way Zeus would modify his beliefs were he to learn this proposition, which he too regards as certainly false. This is a case in which an agent's rationale for what she does depends on her "beliefs about the incredible" and her beliefs about her opponents "beliefs about the incredible."

The flipside of (j) is that Hera has a rationale for drinking the nectar only insofar as she is certain that each of the following subjunctive conditionals is true:

If Zeus were to learn Hj then he would continue to retain his certainty that Hera would still decide to drink the nectar at stage j + 1 even if she were to learn Zj.
In other words, Hera has a rationale for drinking the nectar only if she is entirely convinced, right from the start, that refraining on any given turn would not cause Zeus to alter his views about what she is likely to do on future turns. She must, in effect, regard Zeus' beliefs about her acts as causally independent of her decision. Thus, under conditions of CK, the gods' great knowledge makes it impossible for them to send messages one another about what they will do.

This is not to say that they are unable to communicate at all. By refraining from drink Hera can cause Zeus to alter his opinion about the truth of the CK assumptions. Surprisingly though, sending this message is not in Hera's best interest because it would not cause Zeus to alter his views about her future intentions; he would remain convinced that she would drink on her next turn if she got the chance. Zeus would, of course, be forced to revise his opinions about at least some of the hypotheses that comprise the CK assumptions. For reasons I will not go into here, it turns out that that if the assumptions do hold then Hera must believe that Zeus' would deem her irrational if, contrary to fact, he were to learn that they did not hold. The interesting thing is that this need not prevent Zeus from being able to anticipate Hera's acts. Even if he were to come to learn that she acted irrationally on one turn her would continue to believe that she will act rationally on her next turn. To how this can happen we need to appreciate that irrationality is not an all-or-nothing affair; it comes in gradations. Imagine Zeus thinking about how things would be if he and Hera somehow made it to his penultimate turn. While he is quite sure that this incredible event could only occur if, contrary to what he believes, Hera were irrational, he may still be able to predict what Hera would do on her last turn. Even if Hera were so irrational as to miss the point of the backward induction argument ninety-nine times in a row, he might reason, she would still be so irrational as to pass up a free cup of nectar on her last turn.

To make the idea precise, let's call Hera is grade-1 irrational if she would forgo the nectar on her last turn, when passing it up could not possibly bring her any future benefits, and grade-2 irrational if she would forgo the nectar on her next to last round, in vain hope of future benefits, but would drink on her last round. The point is that if the CK assumptions do hold, then learning that Hera is grade-2 irrational would not lead Zeus to conclude that she is grade-1 irrational, and Hera will recognize this. More generally, if we distinguish one hundred grades of irrationality in Centipede, ranging from (the really serious) grade-1 irrationality up to the (mild) grade-100 irrationality of a person who cannot follow the backward induction reasoning back one hundred steps but can follow it back ninety-nine, then each player will believe that the other would not treat evidence of grade-j irrationality as any evidence of grade-(j + 1) irrationality.

Hera's rationale for drinking the nectar is now obvious. Even though she can convince Zeus that she is irrational by forgoing the nectar, she knows that this would do her no good since he would only deem her mildly (grade-100) irrational and not irrational enough to refrain from drinking on her next turn (grade-99). It makes sense for Zeus to believe this because he is sure that, if given the chance, he could convince Hera that he is irrational by refraining from drink on his first turn, but that this would do him no good since she would only infer that he is grade-100 irrational and not grade-99 irrational. More generally, each player is certain of the following:

If I were to get to my jth turn then the other player would regard me irrational to grade-(j –1) but not to grade-j, and if I were to forgo the nectar on my jth turn the other player would come to regard me as irrational to grade-j but not to grade-(j +1).
This shows us how strong the common knowledge assumptions driving the backward induction argument really are. When Centipede players satisfy these assumptions each will believe that it is impossible to shake the other's confidence in her rationality sufficiently to convince him that she might refrain from drink on her next turn. Under these conditions the only reasonable thing that for the players to do is to drink the nectar the first chance they get, just the backward induction argument predicts.

This is not a conclusion we should resist. We just need to appreciate it for what it is: a conclusion about players who know so much about one another and the game they are plating that they cannot send meaningful signals. That they end up doing poorly in Centipede is not surprising since the game is set up in such a way as to penalize players for having too much knowledge and too much rationality. The only rational strategy when playing such a game is to get out as quickly as possible.

Let me emphasize that this treatment of backward induction assumes that we can make good sense of rational agents having beliefs about what would occur if "incredible" propositions, propositions that they are certain are false, turn out to be true. If you take that away then the rationale I have sketched for Hera and Zeus falls apart. And, it is not just Hera and Zeus who need beliefs of this sort: you and I do as well! Every time we act our reasons for what we do depends on our views about how things would be, and what other people would believe, were we to act differently, and this remains true even when we are absolutely sure that we will not act differently. This is best illustrated by imagining a truncated version of Centipede in which there are only two envelopes in each pile. No one would refrain from drinking the nectar in this case (unless they thought their opponent was moved by considerations of beneficence or fairness). Why would they drink? Simply because, even though they know they will drink and that their opponent knows this, they are sure that refraining from drink would not cause their opponent to change his views about what they would do on their last turn. Beliefs about "the incredible" are thus an unavoidable element of our reasons for acting as we do.

Of course, saying that such beliefs unavoidable tells us nothing about how people might arrive at them or anything about their logical properties. These are important and difficult questions, which I cannot begin to address here. I do, however, encourage readers to pursue the matter further by taking a look at Chapter 7 of my recent book The Foundations of Causal Decision Theory (Cambridge University Press, 1999) where these issues are treated at length. Even though there remains much work to be done in this area, I hope to have convinced you that no entirely adequate account of rational action will be forthcoming until we come to grips with beliefs about the incredible.