

The very idea of rational irrationality

Politics, Philosophy & Economics

1–19

© The Author(s) 2023

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/1470594X231177640

journals.sagepub.com/home/ppe**Spencer Paulson** *Northwestern University, Evanston, IL, USA*

Abstract

I am interested in the “rational irrationality hypothesis” about voter behavior. According to this hypothesis, voters regularly vote for policies that are contrary to their interests because the act of voting for them isn’t. Gathering political information is time-consuming and inconvenient. Doing so is unlikely to lead to positive results since one’s vote is unlikely to be decisive. However, we have preferences over our political beliefs. We like to see ourselves as members of certain groups (e.g. “rugged individualists”) and being part of those groups depends on having certain beliefs (e.g. about welfare spending). Even if a decrease in welfare spending would be bad for me, I might still benefit by believing in and, consequently, voting for a decrease since my vote is unlikely to make a difference but getting to see myself as a rugged individualist will make a noticeable difference to my wellbeing. It is sometimes argued that this hypothesis fails for empirical reasons. I will argue that things are worse: it is conceptually incoherent. I will do so by first showing that it is a rationalizing explanation and then argue that rationalizing explanations must be reflectively stable from the agent’s perspective. The rational irrationality hypothesis is not.

Keywords

political ignorance, motivated reasoning, rational choice theory, rational irrationality, wrong kind of reasons

There is a certain line of thought that has gained momentum in recent years which purports to show that putatively irrational voter behavior makes sense from the perspective of rational choice theory (Caplan, 2001; 2006; Huemer, 2015; Somin, 2013).¹ In a

Corresponding author:

Spencer Paulson, Northwestern University, 1930 Ridge Avenue Unit B304, Evanston, IL 60201, USA.

Email: spencerpaulson2023@u.northwestern.edu

nutshell, the idea is that a single vote is so unlikely to make a difference to the outcome of an election that the benefits of believing in and, consequently, voting for a disastrous policy will often outweigh the costs. If the agent has preferences over certain beliefs as well as material outcomes, there will be cases where they can maximize utility by trading one preference off for the other. If an agent prefers to believe that climate change is a hoax, they can believe as they wish at no cost to themselves.² They might act on this belief in a way that results in additional pollution, but the amount will be so small that it won't cause any negative outcomes that wouldn't have occurred anyway. The added pollution is not even noticeable, whereas believing that climate change is a sham might significantly increase their sense of wellbeing for whatever reason. Similarly, a policy that decreases environmental regulation might harm one significantly. However, casting a single vote in favor of it is unlikely to determine the fate of the policy. So, if one prefers to be a member of the political group that opposes environmental regulation, the benefits of believing falsely that environmental regulation does more harm than good could very well outweigh the costs.³ In short, the rational irrationality hypothesis explains why people vote against their own self-interest by positing that they have preferences over beliefs. As a result, they sometimes maximize utility by believing something false about which policies would benefit them. These false beliefs lead them to vote for policies that are contrary to their own self-interest. Their votes are unlikely to be decisive, so the votes have no negative consequences.

Call the phenomenon just described "rational irrationality." The theory of rational irrationality is a species of rationalizing explanation. It explains why agents do something by showing that it makes sense for them to do it. Given the circumstances and the agent's preferences, it makes sense for them to have false beliefs about certain policies and this leads them to vote for disastrous policies. They so believe because they are implicitly aware of this. This is obscured by the fact that the target behavior is simultaneously described as being irrational, in some sense. However, I will argue that this is misleading. Proponents of the rational irrationality hypothesis are not always careful about how they state their position. When they are, they are clear that instances of "rational irrationality" are indeed rational in the only way that matters (though see the objections and replies for a qualification).

This last fact is important because it commits proponents of the rational irrationality hypothesis to the intelligibility of rational irrationality from the point of view of the agent. This is not always clear because proponents will sometimes ward off criticism by granting that rationally irrational thinking can only occur subconsciously. Nobody can be rationally irrational and also be fully aware that this is what they're doing. However, one of the burdens of this article will be to show that this cannot (given their other commitments) mean that rational irrationality is conceptually incoherent or that a fully rational agent would be incapable of it. Rather, since their explanation of behavior requires that "rational irrationality" really is rational in the sense that matters, a fully rational agent would not only be capable of it but, if they had the right preferences, would be rationally obliged to engage in it. Any inability of ours to self-consciously partake in the "rationally irrational" must on their account be a merely contingent (and unfortunate) psychological limitation.

I will argue that commitment to the intelligibility of rational irrationality is a problem. Intuitively, the practice seems to be reflectively unstable. I will vindicate this intuition,

showing why it is. I will do this by drawing a parallel between rational irrationality and Hieronymi's (2005, 2006, 2009) solution to Kavka's (1983) "toxin puzzle." Both involve cases where it at first seems that it would make sense to form a mental state that commits one to doing something even though it would not make sense to do that thing. I will argue that in both cases the mental state is irrational despite its favorable consequences, because it rationally commits the agent to a course of action that is contrary to her own self-interest.⁴ Just as intending to Φ commits one to Φ -ing, believing that a policy is best all-things considered rationally commits one to that policy. It follows that one cannot treat the question of what to do about a political issue and what to believe about it as two different questions, admitting two different cost-benefit analyses, as the rational irrationality view requires. This criticism, if successful, is more devastating than extant criticisms which hold that the rational irrationality hypothesis is mistaken on empirical grounds.⁵ If I succeed here, I will show that the very idea of rational irrationality is conceptually incoherent.

In the Rational irrationality section, I will explain the main tenets of and motivation for the rational irrationality hypothesis. In the Rational irrationality and rationalizing explanations section, I will argue that the style of explanation that this view involves is a rationalizing one and that this commits its proponents to the rational intelligibility of it from the agent's point of view. In the Rational irrationality and reflective stability section, I will draw a comparison between the toxin puzzle and rational irrationality. Here I offer an explanation of why the position strikes us as reflectively unstable. In the Objections, replies, and follow-up discussion section, I will consider an objection to the effect that we can save the rational irrationality hypothesis by distinguishing epistemic and practical rationality. I will also briefly consider the form that explanations of motivated reasoning must take once the rational irrationality hypothesis has been ruled out.

Rational irrationality

I will begin with a brief overview of the rational irrationality hypothesis. The impetus was the apparent difficulty rational choice theory faces when trying to explain putatively irrational behavior. Few of the formal details of rational choice theory matter for my purposes. I will describe this approach to modeling and/or explaining behavior at a level of abstraction that omits a great deal.

Rational choice theory is a way of modeling or explaining (more on the distinction below) the behavior of agents. Agents have preferences. Preferences are given a total ordering. They are ordered by utility to the agent. Utilities are given quantitative values. The agent aims to maximize utility. To do this, they must perform cost-benefit analyses of the courses of action they consider. Some outcomes have negative utility, that is, costs. Others have positive utility, that is, benefits. Agents must weigh the costs and benefits of all the outcomes for each course of action. They choose the course of action with the most favorable balance, that is, the rational one. Some versions of the theory assume that the agent has all the relevant information. Others have provisions for uncertainty. On views of the latter sort, agents maximize expected utility, as opposed to utility *simpliciter*. Expected utility is just the utility of an outcome multiplied by the probability that it obtains.

Economists, behavioral economists, and social psychologists often resort to rational choice theory for the purpose of making predictions about the market.⁶ They assume that everyone is behaving rationally (in the above sense).⁷ Monetary value is the only thing with any utility in the simplest models. This makes it easier to generate predictions since it is easy to give monetary preferences quantitative values and put them in a total order. However, the approach can be extended to agents with more nuanced preferences, so long as they can be given quantitative values and put in a total order.

The theory runs into apparent difficulties when agents seem to behave irrationally. It is sometimes thought that voters tend to behave irrationally. This thought has been around at least since Plato,⁸ but it picked up steam in the 20th century in part because of Joseph Schumpeter's (1950) *Capitalism, Socialism and Democracy*. If voters don't behave rationally, then rational choice theory isn't going to help us predict or explain their behavior.

Anthony Downs (1957) argued that things aren't as bad as they might seem for rational choice theory. Voters vote for policies that are contrary to their own interests because they are ignorant. However, it is rational for them to be ignorant. The time spent pursuing political knowledge has negative utility. The negative utility is not compensated by the positive utility of voting for sound governmental policies. Perhaps it would be if one's vote were decisive. However, one's vote is unlikely to be decisive, even in a local election. Let the expected utility of an act be the sum of the products of its possible consequences and their respective utilities. An agent does not maximize expected utility by casting an informed vote because the negative utility of acquiring information is moderate (let's suppose) and nearly certain to obtain whereas the scenario in which one's vote is decisive has high utility, but nearly certain not to obtain. The intuitive thought here is that the expected utility of voting is low because it is unlikely that one's vote makes a difference, but it is very likely that becoming informed will be onerous and involve additional opportunity costs. That is, voters are ignorant but rationally so. You maximize expected utility by remaining politically ignorant, unless you have a preference for political knowledge as such. Assuming such preferences are rare, or at least weighted below more pressing concerns, rationality requires that one remain ignorant. Call this phenomenon "rational ignorance."

So, rational choice theory predicts that voters will be poorly informed because it predicts cases of rational ignorance. That means it makes sense that they vote for policies that are contrary to their own interests. Or so it might seem. However, Bryan Caplan (2001, 2006) has argued that rational ignorance fails to explain the conviction with which people favor policies contrary to their own interests. If one is ignorant, one should be agnostic (Caplan, 2006: Introduction, Chapter 1). However, we find something closer to fanaticism than agnosticism among the uninformed.

Does this pose problems for rational choice-theoretic approaches to voter behavior? Caplan thinks not. The solution is to treat beliefs as one more thing over which agents have preferences (Caplan, 2001:7ff, 2006: 115ff). People prefer to believe certain things, for example, that God exists or that immigrants are destroying the economy. These beliefs have positive utility for them. Even if these beliefs are false, their falsity does not bear on the decision making of the agent in such a way as to incur negative-utility outcomes, as many false beliefs clearly do. Having a false belief about

which treatment plans are likely to cure a disease could have terrible consequences for the agent. Having false beliefs about immigration policy is much less likely to have noticeable consequences for the agent. The agent can enjoy these beliefs and vote for policies that would have negative utility were they enacted. However, since the agent's vote is unlikely to be decisive, this doesn't make much of a difference from the perspective of rational choice. This is so because the inquiry would itself be costly, as would the loss of a cherished belief. Furthermore, no positive utility would come of the true belief unless the agent's vote was decisive. So, instead of "rational ignorance," we now have "rational irrationality."

This approach has been criticized in many ways. It has been argued that the empirical studies Caplan cites falsely assume that "textbook" knowledge (e.g. knowledge of how many senators each state has) is necessary for reasonable democratic participation (Landmore, 2012). There are worries that those who conducted those studies framed the questions in such a way as to encourage participants to give the wrong answer (Landmore, 2012).⁹ Caplan's assumption that votes only matter if they are pivotal has also been challenged (Mackie, 2012). My criticism will be more damning than any of these. Rather than arguing that Caplan is empirically mistaken or that one of his premises is incorrect, I will argue that his account is incoherent. To do that, I will first need to go over some general features of the account he is offering. I will focus on Caplan's account here because it is the most rigorous version of the rational irrationality hypothesis I've come across. However, I will occasionally draw from the work of other similarly minded proponents of rational irrationality. The conclusion I reach in this article will apply to every version of the rational irrationality hypothesis since it depends only on its most general features.

Rational irrationality and rationalizing explanations

As I said earlier, rational choice theory can be pursued as a genuine explanation of behavior. A genuine explanation is an answer to the question why the behavior occurred. It can also be treated as a mere heuristic. If one takes the latter route, one assumes that agents will behave rationally just to make predictions about their behavior but will remain undecided as to whether rationality really has anything to do with why they behave as they do. Rational choice theory could be a useful way of generating predictions about behavior even if it doesn't give us a true explanation of why it occurred.

Rational irrationality, then, could be taken as either genuine explanation or mere heuristic. Which way does Caplan understand it? One consideration that may seem to speak in favor of the heuristic interpretation is Caplan's concession of the lack of "psychological plausibility" of the rational ignorance proposal, at least understood a certain way. He grants that people don't explicitly reason this way.¹⁰ However, he goes on to say that people do tacitly make the trade-offs he is attributing to them. They tacitly recognize that learning more is not worth the effort and that holding certain beliefs feels good,¹¹ so they tacitly lower their intellectual standards when it enables them to hold low-cost feel-good beliefs.

How should we understand "tacit" awareness? Caplan suggests we understand it by way of an analogy with driving a car (Caplan, 2006: 126). We don't explicitly reason

about what we are doing but “we know the steps on some level” (Caplan, 2006: 126). My knowledge of how to drive a car is presumably psychologically real in some important sense, despite not being explicit. I don’t simply act as if I know how to drive a car. I really know how to drive a car. It is an interesting question what this knowledge consists in and how it is realized. Is it discursive? Is it brain-bound? If so, is it local or distributed? Regardless of how we settle these questions, we should grant that this knowledge is real and that it is explanatorily relevant at a certain level of description. I hit the brakes *because* that is how you stop a moving car and I know this.

Caplan goes on to try to demonstrate the psychological plausibility of rational irrationality, understood as a subconscious process. He provides numerous examples of cases in which agents put their rationality “on standby” when they can hold high-utility false beliefs at low cost to themselves, but then revise those beliefs when the stakes go up and those beliefs cease to come at a low cost (Caplan, 2006: 127–131). He is explaining their behavior by positing that it is the result of a tacit decision to engage one’s rational capacities that were previously on standby. They were on standby because cost-benefit analysis deemed that the best course of action. As Caplan puts it, “Voter irrationality is precisely what economic theory implies once we adopt introspectively plausible assumptions about human motivation” (Caplan, 2006: 3, cf. 14). That is, he is not only predicting how people will behave but also giving an account of the factors that motivate them to so behave. Huemer’s (2015) paper on the topic, which cites Caplan with approval and argues for the same conclusion, is called “Why We are Irrational about Politics.” This strongly suggests that he is interested in explaining why we do the things we do and not just giving us a behavioristic ersatz.

So, we are dealing with a genuine attempt at explanation rather than mere prediction. Now we must determine what kind of explanation he is offering. It seems clear that he is offering a rationalizing explanation. Rational choice theory explains behavior as happening because it is rational for the agent.

The question we must consider now is whether his explanation of “rational irrationality” is one that depends on the subject being mistaken. Of course, in one sense the answer is “yes.” He is assuming that voters are wrong about which policies make sense and he is trying to explain why they are wrong. The important question however is whether the behavior he is explaining would still make sense from the point of view of the agent even if they were aware of what they were doing. Is Caplan saying that the rationally irrational are acting on merely apparent reasons that do not appear as reasons to act from the point of view of the enlightened theorist providing the explanation? Or is he saying that even when all the facts are in, it is rational to be rationally irrational, so long as you have the right preferences? Another way of putting the question is to ask whether Caplan is offering an error-theory. If yes, then the rationally irrational only behave this way because they are in the dark. If no, they would continue to behave this way even if they accepted Caplan’s theory.

The error-theoretic answer has *prima facie* plausibility. Caplan’s reader gets the distinct impression he takes himself to be smarter than the average voter and that he is debunking their follies with empirical facts and explaining why their mistakes persist. They do as they do because they are willfully irrational and they would cease to be so if they were to soberly reflect, as he has done.

I have little doubt that Caplan is tempted, in certain moods, to see matters this way. However, I will argue that he is not really an error-theorist. I hope to show that his preferred method of explanation requires him not to be. When he is being careful, he seems to appreciate this fact. I now turn to make the case that he is not offering an error-theoretic explanation and that he only gives this impression when he is not being careful.

I will begin by considering grounds that one might think speak decisively in favor of the error-theoretic interpretation of Caplan. During most of his discussion of the psychological plausibility of his account, he takes for granted that the rationally irrational cannot be self-consciously so. There is a temptation here to infer that this must be because rational irrationality is a reflectively unstable position. One can benefit from rational irrationality so long as it takes place in the Freudian depths. One cannot, however, consider the merits of rational irrationality and decide in favor of it on that basis. Rational irrationality is a species of irrationality. One cannot deliberately be irrational, just as one cannot deliberately fail. One can deliberately succeed at something other than what one is putatively trying to do, such as when one deliberately loses a sporting event. However, one cannot genuinely engage in a task and deliberately fail to achieve its constitutive aim.¹² If one is trying to fail at the constitutive aim of one pursuit, one is really engaged in some other pursuit. If one is trying to lose a baseball game, one is “playing” baseball in a pickwickian sense only. What one is really doing is profiting by gambling on a fixed game. Thought is governed by a rationality norm, or so one might think. Plausibly, believing something is taking it to be true. Belief in some sense aims at truth,¹³ then. If so, then rationality is at least a secondary norm of belief.¹⁴ Belief aims at truth and rationality is, plausibly, at least a guide to truth. So, you ought to only believe that which it is rational to believe.¹⁵ The truth and rationality norms are essential to belief just as the aim of scoring more runs is essential to baseball. So, one cannot think and deliberately do so irrationally. Therefore, explanations of behavior invoking rational irrationality are error-theoretic. An agent might receive certain benefits from being rationally irrational, but they cannot decide to be rationally irrational in order to secure these benefits.

This is a very reasonable line of thought. I now turn to show that it is not what Caplan thinks and, furthermore, even if he did it is inconsistent with his other commitments. Initial grounds for doubt appear when he suggests that people really can be rationally irrational willingly and knowingly. Caplan quotes from *1984* a passage in which those partaking in “doublethink” are said to engage in contradictory beliefs simultaneously and self-consciously (see Caplan 2006: 125–126). In one act of doublethink, one self-consciously violates the requirements of rationality and follows it with another act in which one’s memory of the previous act is erased. Memory is erased to mitigate the feeling of guilt (Caplan, 2006: 125–126). Caplan doesn’t discuss the Orwell quote at-length because he doesn’t think his case depends on “Orwellian underpinnings” (Caplan, 2006: 125). However, the fact that he thinks the Orwell passage gives a “chilling” (Caplan, 2006: 125) account of a real phenomenon suggests he doesn’t think that one cannot engage in rational irrationality knowingly. He will accept for the sake of argument that doublethink doesn’t really happen, but he is only doing this because he doesn’t expect people to agree with him that it does. So, he is arguing from premises that his reader is more likely to accept. If you can engage in doublethink knowingly and

deliberately, then it doesn't seem that the line of thought adumbrated above is one Caplan endorses. However, there are additional reasons more significant to his overall project for reaching this same conclusion.

Although Caplan is less clear than he could have been about this, I am persuaded that he thinks self-conscious doublethink is possible. He thinks this because he does not believe that rationality, in the sense of "rationality" that is being "put on standby" in cases of rational irrationality, is a particularly important normative standard, let alone constitutive of thought. The rational choice theorist's sense of "rationality" is the only one he takes to be authoritative in his more careful moments. It is worth noticing in this regard that he sometimes puts "irrationality" in scare-quotes when he talks about rational irrationality (see, e.g. Caplan 2001: 4). This strongly suggests he is not endorsing the negatively valenced evaluation that the term carries with it. He does not put "rational" in scare-quotes.

The sense of "rational" in which rational irrationality is irrational is that of responsiveness to the evidence. Most, although not all, philosophers are inclined to think that epistemic rationality is not simply a special case of practical rationality. Epistemic rationality, rather, is just as fundamental as practical rationality, that is, the kind of rationality that the rational choice theorist hopes to illuminate. Some such philosophers, "evidentialists" maintain that a distinctively epistemic kind of rationality is the only standard relative to which we can assess the rationality of beliefs (e.g. Moran, 1988; Shah, 2004). Others maintain that there are two incommensurable standards relative to which we can assess the rationality of a belief (e.g. Feldman, 2000). We can assess it relative to the purely epistemic standard that evidentialists say is the only intelligible standard for belief rationality, but we can also assess the practical rationality of beliefs. The latter standard may be analyzed in terms of expected utility.¹⁶ Either way, the majority say there is a purely epistemic standard relative to which the rationality of beliefs either can or must be assessed.¹⁷

Caplan does not share this view. On his view, beliefs are just another thing we have preferences over. Unless you assign an unusually high utility to true beliefs as such or in some specific domain, there will be cases where it is rational to make trade-offs. When the costs of getting things right exceed the benefits of doing so, it is irrational to take the necessary measures to get things right. The like-minded Ilya Somin clearly makes this point when he says of rational irrationality that "Although some scholars view such bias as irrational behavior, it is perfectly rational if the goal is not to get at the 'truth' of a given issue in order to be a better voter but to enjoy the psychic benefits of political 'fun'" (Somin, 2013: 80). He then goes on to cite Caplan with approval. It seems clear that the standard he is using to assess the rationality of beliefs is the same standard he would use to assess any other behavior, such as buying a lottery ticket.

Caplan says similar things in many places. He describes cases of "rational irrationality" as ones in which people have preferences over both beliefs and material goods and "rationally make trade-offs between their two values" (Caplan, 2006: 17). He clearly does not regard epistemic rationality as something on a par with practical rationality. Nor does he think the two are incommensurable, for if they were, they would be incomparable and trade-offs between them could never be rational. Rather, just like Somin, he takes epistemic rationality to be what you get when you only assign utility to true beliefs.¹⁸ Epistemic rationality is practical rationality with a peculiar utility function. Believing in accordance with the evidence no matter the practical consequences is only rational

if you have the sole goal of achieving true beliefs. But that would be a niche interest of yours, not a feature of rationality as such. So long as you value something other than true beliefs, there are cases where rationality requires that you make trade-offs. The real problem with false beliefs is that there are material benefits you sacrifice on account of them (Caplan, 2006: 17). However, sometimes we sacrifice little in the way of material benefits and we enjoy the false beliefs enough to compensate for the loss. So, we would be irrational not to make the trade-off in such cases.

This is why he puts “irrationality” in scare-quotes when he’s careful. The trade-offs are rational by the only standard he takes seriously. The non-economist’s intuitions about rationality are just a special case of rationality as understood by rational choice theory. It is what rational choice theory gives you when you plug in the utility function of an epistemic fanatic. There is nothing wrong with being a fanatic, but there is nothing wrong with not being one either, from the point of view of rational choice theory. Most people aren’t fanatics. So, it is rational for them to make the trade-offs.¹⁹ If he is offering an error-theory of anything, it is the deep-seated conviction that there is something wrong with these trade-offs for agents who maximize expected utility by making them.

Caplan is committed to saying that our inability to engage knowingly in rational irrationality is a mere psychological quirk of ours. A fully rational agent would be knowingly rationally irrational if they had the right preferences. We are unfortunately unable to consciously make the trade-offs that rationality requires of us, so we must make the trade-offs subconsciously. Caplan’s explanation does not require that the agent be unaware of what they’re doing. They could, at least in principle, be aware. The agent’s goal is to maximize utility and the way to do that is by subjecting certain beliefs to a lower level of scrutiny. So, they will do so if they accept Caplan’s theory. They must, on pain of irrationality. To do anything else would be to fail to maximize expected utility. The only way this would fail to be the case, on Caplan’s view, is if they were to be an epistemic fanatic. So, saying that agents are putting their rationality “on standby” in these cases is misleading. They are just ceasing to pursue one of their subsidiary aims when it is no longer profitable to do so.

This puts Caplan in an awkward position. He seems to be lecturing the *hoi polloi* while at the same time saying that they are doing the only thing it makes sense for someone in their position to do. This should not distract us from the fact that his explanation of voter behavior, which is his main project, depends on the rationality of their choices. Just as Downs explains voter ignorance by showing that it makes sense for the agent to be ignorant, all things considered, so too does Caplan explain voter “irrationality” by showing that the putative irrationality is, contrary to appearances, rational. It follows that rational irrationality ought to be reflectively stable. If it is what rationality requires, it is what a fully rational agent would self-consciously decide to do. Our inability to do it self-consciously is a failure on our part to be self-consciously rational. I now turn to show that this is a problem for Caplan and his followers.

Rational irrationality and reflective stability

The difficulty with rational irrationality is that it presents us with an apparent example of a case where it is rational to form a psychological attitude in favor of doing something but

not rational to do it. It sometimes benefits you to believe a policy is beneficial even though the policy itself would not benefit you. It seems to follow that it makes sense for you to believe it is best all-things considered to have a policy enacted although it would not make sense for you to enact the policy, if the decision were up to you. Surely, this is the feature of rational irrationality that is responsible for our uneasiness about it. This is where the irrational part of “rational irrationality” comes in. If it is reflectively unstable, it must have something to do with this feature. However, this falls short of explaining its reflective instability. It only gestures in the direction of an explanation.

We can make progress toward an explanation by considering what people have said about similar phenomena. So, let us consider another putative case where it is rational to decide in favor of doing something but not rational to do it. Take Pamela Hieronymi’s (2009)²⁰ variation of Kavka’s (1983) “toxin puzzle.” Suppose a scientist has a reliable intention-detector. She is offering you \$100 to intend to jump out of a third-storey window. She is willing to pay you the \$100 just for intending; you don’t need to jump. She will pay the money if the intention comes to fruition and you jump, although you needn’t to collect. Seemingly it would be worth just intending to jump to collect the money even though it wouldn’t be worth jumping. The intention itself won’t injure you. It seems, then, that you have sufficient reason to intend to jump although you do not have sufficient reason to jump.

Most of us get the sense that something has gone wrong here, although it is hard to say what. Niko Kolodny has offered the following straightforward solution.²¹ It isn’t rational to intend to jump because intentions to jump tend to result in jumping and it isn’t rational to jump. If this is the solution to the problem, then there is a sharp disanalogy between Hieronymi’s case and rational irrationality. The probability of me discharging an intention conditional on my having that intention is very high. I don’t discharge all my intentions, but I discharge most of them. However, the probability of me deciding the outcome of an election (or referendum) conditional on my forming a political belief is very low. The belief is likely to produce a vote, but the vote is unlikely to decide the election. So, rational irrationality is rational although taking Hieronymi’s scientist’s offer is not.

I worry though that Kolodny’s answer doesn’t get to the heart of the matter. It seems that the probability of me discharging the commitment is downstream from what really matters for the purposes of rationality. Kolodny’s view seems to involve us treating ourselves as machines that set themselves in motion and then wait to see what happens. We decide on our ends by regarding ourselves empirically and determining what is likely to happen if we pursue those ends. We might reasonably decide to pursue ends we regard as awful if we are pretty sure our pursuit of them will fail but lead to desirable consequences along the way. That sounds odd, we should look for a better answer.

Even if some sort of empirical awareness of ourselves is required for agency, it does not follow that this is our primary mode of self-awareness in practical deliberation, nor must we resort to an empirical way of knowing ourselves to see what is wrong with taking the scientist’s offer. I won’t go into the details of Hieronymi’s account, but the key aspect of it is that what it is to form an intention to ϕ is to affirmatively settle the question whether to ϕ or, alternatively, to commit oneself to ϕ -ing. Of course, one might affirmatively settle the question whether to ϕ without ever ϕ -ing, either because of unforeseen circumstances or akrasia. Nonetheless, when one forms an intention to

φ , one is committed to φ -ing. One can form an intention to φ and be akratic, but one cannot do this on purpose. If one tries to do this, one fails to settle the question of whether to φ affirmatively and therefore does not really intend to φ .

It is unreasonable to take the scientist's offer because once one has formed the intention to jump, one has committed oneself to a course of action that is irrational by one's own lights. Jumping is not worth \$100. Yet one is committed to jumping, since intending to jump is a commitment to jumping. The main problem with forming an intention to do something foolish is not that it leads to foolish behavior down the road, as combining alcohol and caffeine might. The problem is that once one has formed the intention to do something foolish, one is already rationally committed to performing the foolish act and hence the proper subject of rational criticism for it, even if the intention never comes to fruition. The reasons for intending to do something, consequently, do not come apart from the reasons for doing it. That is, as is sometimes said, the reasons relevant to whether to intend to φ are "transparent" to the reasons for φ -ing (cf. Evans, 1982; Moran, 2001). Intending to jump and jumping don't receive separate cost-benefit analyses,²² despite the fact that intending to do something and actually doing it are not the same thing.

I propose that the best explanation of our uneasiness about rational irrationality is that the same thing is going on. Reasons for believing a policy is all-things-considered best are transparent to reasons for adopting that policy. In believing a policy is all-things-considered best, one is rationally committed to it. It doesn't matter if one's belief is unlikely to cause a decisive vote, just as rational criticism of the agent who takes the scientist's offer does not require us to take note of how likely it is that they discharge their intention. Note that in order to secure the benefits of rational irrationality, one has to be genuinely committed to the policy that is (in fact) contrary to your best interest. The rationally irrational subject wants to (for example) be the kind of person who is genuinely committed to a decrease in welfare spending: they want to be the kind of person who genuinely believes it is best to decrease it. This is why rational irrationality is like the toxin puzzle. In both cases, the subject is rationally committed to something contrary to her own self-interest. The commitment itself is conducive to positive utility (and no negative utility), but the commitment is nonetheless irrational.

Of course, Caplan will have the following rejoinder. He is not saying that people first figure out that policies are bad or that beliefs are false, and then form the belief anyway (Caplan, 2006: 126). Rather, one is tacitly sensitive to the costs and benefits of continued inquiry and tacitly decides against carefully considering certain matters because the reward is not worth the effort, the time, and the potential loss of cherished beliefs. But one does not first discover that the cherished beliefs are false. So, it is not quite like the toxin puzzle where the agent already knows what happens if they jump.

Caplan is free to say that this is not what human reasoners in fact do. However, I hope to have shown that he is committed to "rational irrationality" being rational in the only way that matters. This means that, regardless of what we human reasoners in fact do, it would be perfectly rational to explicitly go through the steps we go through tacitly. That's a problem because rational irrationality is reflectively unstable, not because of a peculiarity of human psychology, but rather because it is incoherent.

On Caplan's account, people don't vote "irrationally" because they are mistaken about the requirements of rationality or the empirical facts relevant to their decision. They are

aware, however tacitly, of what the likelihood of casting a decisive vote is, what their preferences are and what the costs of inquiry are. On Caplan's rational choice-theoretic account of rationality, these are the only factors relevant to the agent's decision. The agent is not mistaken about any of them. So, if one were determined by the force of the better reason and were aware of all the relevant facts, one would explicitly know exactly what the human reasoner implicitly knows, and they would make the same decision about what to do on the same basis.

Suppose the fully rational subject is aware that their cherished beliefs are false. It is not clear why this matters if Caplan is right. The rational subject does what the canons of rationality require. If Caplan is right, they require only that one maximize expected utility. So, if they have the same preferences as the rationally irrational human reasoner, they will do the same thing, but explicitly. That is, they will believe the falsehood, vote on its basis, and so forth. Maximizing expected utility only requires following the evidence if one has a preference for true beliefs that is weighted so that trade-offs don't make sense. From the point of view of rationality as such (as Caplan understands it), there is no reason to have one set of preferences rather than another. Reason is just a slave to the preferences.

So, even if a reasoner were to do exactly what Caplan believes rationality requires of them and they were to be supplied with the information that a policy they favor would be disastrous (in terms of utility) if implemented, they would still have to treat the question of whether to believe it is a good policy as a separate question. The fact that the policy would be disastrous if implemented only weighs on their deliberations about what to believe to the extent that their belief is likely to cause the policy to be adopted.

This makes it much like the case in which an agent treats the question of whether to form an intention to jump as orthogonal to the question of whether to jump except insofar as the intention is likely to come to fruition. The problem with the view of rationality and agency espoused by Caplan and suggested by Kolodny's brief remarks is that commitments are treated as mere instruments rather than themselves the locus of rational assessment. Commitments are themselves the locus of rational assessment and intending to ϕ is a commitment to ϕ -ing, so rationality does not require one to undertake a commitment to do something foolish even if the commitment itself has good consequences. One would still be the proper subject of rational criticism. Belief is no less commitment involving than intention (more on this in the next section). So, for the same reasons, forming a belief that a silly policy is all-things-considered best is to be committed to it. This makes one the proper subject of rational criticism, even if one's vote is not decisive.

Objections, replies, and follow-up discussion

The first objection to be considered is that the rational irrationality hypothesis can be saved if we distinguish between epistemic and practical rationality. Caplan himself didn't do this, but others such as Huemer (2015) have. Perhaps believing that a policy is beneficial is epistemically irrational but practically rational. We might think that this is what rational irrationality consists in: a belief that is rational according to one legitimate standard of rationality and irrational according to another, perhaps incommensurable standard.²³

My response is to grant that epistemic rationality and practical rationality are distinct, possibly incommensurable, standards and that they can come apart, but they can't come apart in the way the proponent of the rational irrationality hypothesis needs them to come apart. To see how they can come apart, consider a case where the subject has a serious disease, and the probability of recovery is very low (cf. Feldman 2000). However, the probability of recovery increases significantly if the subject believes that they will survive. In this case, it is epistemically irrational for the subject to believe that she will survive. The evidence tells against it. However, it is practically rational for her to believe that she will survive, since forming that belief is instrumental to a better outcome than can be secured by believing in accordance with the evidence in this case. Here is a case where the practical and the epistemic clash.

We might think that cases of rational irrationality work the same way. After all, the subject believing something against her evidence is conducive to greater personal well-being when the false belief that will result makes her feel good and doesn't alter the result of the election. However, I urge that this case is different in an important respect than the case of the serious disease. The belief in that case was not about what it is best to do. However, all-things-considered beliefs about what to do carry along practical commitments that bear on practical rationality. In particular, if you believe it is best all-things-considered to Φ , then you are rationally committed to Φ -ing.

To see why, consider the following variation of the toxin puzzle: the scientist is offering you \$100 to believe that it is all-things-considered best to jump out of the window. This seems like essentially the same problem as the one considered earlier. It isn't rational to form the belief, even though there is no harm in just believing any more than there is in just intending. The reason is that the belief rationally commits you to jumping out the window, just as the intention to jump does. If you judge that it is all-things-considered best to Φ , then you are akratic if you don't try to Φ . This shows that the belief itself brings with it a practical commitment. So, the belief is practically irrational despite being instrumental to maximizing utility. It is practically irrational because it commits one to a course of action that is itself irrational by one's own lights. Recall from the last section that the rational irrationality hypothesis doesn't turn on the subject being unaware of any empirical facts and the policy is in fact contrary to the subject's best interests.

One further complication needs to be addressed before moving on. The case of believing a policy is best-all-things-considered is a bit different than the case of believing that it is best all-things-considered for the agent to do something. Enacting a policy is something a group of which the agent is a part does rather than an action performed by the agent herself. However, it isn't clear how this helps. Consider the following variation of the toxin puzzle: the scientist offers you \$100 to believe that it is best-all-things-considered for us to enact a policy that will result in our defenestration. I urge that this is problematic in much the same way as the original toxin puzzle.

If you are on board with what I have said so far, then you might well wonder what we should say about alleged cases of rational irrationality.²⁴ There surely are cases of people voting against their own best interest and doing so because they have mistaken beliefs about which policies would be in their own best interest. What are we to say about them?

I cannot offer a full explanation here. I will only say enough to explain how we avoid the difficulties I have raised for the rational irrationality hypothesis. I argued in the

Rational irrationality and rationalizing explanations section that the rational irrationality hypothesis is essentially a rationalizing explanation as opposed to an error-theory. That is, the subject does as she does because it is rational for her to do it, not because she is mistaken about what is really in her own self-interest. The upshot of my argument is that we need to resort to an error-theory instead because it is not rational (practically or otherwise) for her to believe as she does. The error-theory can presumably incorporate many of the points Caplan and others make about the costs of inquiry outweighing the benefits. This can help explain why the subject is mistaken about her own self-interest without committing us to the problematic claim that it is rational for her to be so mistaken. Her mistaken belief is now seen as the unfortunate byproduct of an otherwise rational course of action rather than an end in itself.

Conclusion

I have argued that rational irrationality is incoherent, and this is a problem for those who wish to invoke it as a rationalizing explanation of behavior. I have argued that Caplan and company are offering the rational irrationality hypothesis as a rationalizing explanation of voter behavior. So, their explanation fails. The correct explanation will be an error-theory, rather than a rationalizing explanation. For this reason, rational choice theory is ill-equipped to explain the phenomenon of people voting contrary to their own self-interest.

Acknowledgements

I would like to thank Cristina Lafont and Sandy Goldberg for helpful comments on earlier drafts of this article and two anonymous referees for helping me develop the strongest version of the argument.


Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Spencer Paulson  <https://orcid.org/0000-0001-5794-1387>

Notes

1. A similar approach to explaining factual polarization and the clustering of logically independent political convictions is also prevalent (see Cohen, 2003; Greene, 2013; Jost et al., 2013; Kahan, 2016a, 2016b; Sherman and Cohen, 2006).

2. Proponents of rational irrationality generally do not assume doxastic voluntarism. More on this below.
3. Cf. Kahan (2016a).
4. Throughout the article I will focus on cases of voting against one's own self-interest. I do this in part for ease of exposition and in part because behavioral economists often do the same. The claim that voting against your own self-interest is irrational might appear to commit me to egoism. It doesn't. In the cases of interest to behavioral economists, the voter is *unintentionally* voting against her own self-interest. That is, she is trying to vote for policies that benefit her, but she is confused about which ones those are. Furthermore, the confusion is generated by motivated reasoning rather than misleading evidence. Perhaps sometimes voting against your own self-interest is rational, but presumably not in these cases. Thanks to an anonymous referee for bringing this issue to my attention.
5. See Landemore (2012) and Mackie (2012).
6. For influential applications see Coleman (1964), Downs (1957), Homans (1958), Olson (1965), and Schelling (1960).
7. Unsurprisingly, this approach has its shortcomings due to this assumption. For some well-known ones, see Ariely et al. (2003), Kahnemann and Tversky (1982), and Madrian and Shea (2001). These are orthogonal to the issues I will be pursuing here.
8. The *locus classicus* of this line of thought being his (2004) *Republic*.
9. Michael Hannon (2022) argues that more knowledgeable voters tend to be more biased, which is supported by some empirical findings (e.g. Kahan, 2013). This might complicate things for Caplan, since Caplan is trying to explain puzzling voter behavior in terms of lack of knowledge and then explain lack of knowledge in terms of rational irrationality.
10. Although this might just be for the sake of argument, more on this below.
11. Cf. Somin (2013: 80).
12. Brewer (2009) makes the same point.
13. Cf. Adler and Hicks (2013), Boghossian (2003), Engel (2005, 2007, 2013), Humberstone 1992, Millar (2004), Owens (2003), Shah (2003), Shah and Velleman (2005), Steglich-Peterson (2006, 2009), Vahid (2006), Velleman (2000), Wedgwood (2002; 2013), Whiting (2012; 2013), Williams (1973: 136–137), and Zalabardo (2010). This claim seems plausible, though it has been challenged (see Bykvist and Hattiangadi, 2007, 2013; Gluer and Wikforss, 2009; Horwich, 2013; Papineau, 2013). If you disagree with me on this, you can still agree with my conclusion in this article. In fact, you might not even be tempted to read the above into Caplan's position, which makes my job easier, since the purpose of this portion of the article is to describe a position many people assume Caplan endorses in order to then argue that it isn't what he thinks (see below).
14. See DeRose (2002) for more on the primary/secondary distinction. I say "at least" since I leave room for the possibility that the rationality norm isn't derived from the truth norm (see Berker, 2013; Gibbons, 2013).
15. Cf. Wedgwood (2002).
16. I will raise doubts about the project of analyzing practical rationality in terms of expected utility in the objections and replies section.
17. For views that deny this, see Reisner (2007, 2009, 2013) and Rinnard (2017, 2019).
18. This is similar to the approach taken by epistemic utility theorists such as Easwaran (2013), Easwaran and Fitelson (2015), Leitgeb and Pettigrew (2010), and Joyce (1998) although Caplan is talking about full belief rather than credence.

19. Cf. “There may be some people for whom *being epistemically rational* is itself a sufficiently great value to outweigh any other preferences they may have with regard to their beliefs. Such people would continue to be epistemically rational, even about political issues. But there is no reason to expect that everyone would have this sort of preference structure. To explain why some would adopt irrational political beliefs, we need only suppose that some individuals’ non-epistemic belief preferences are stronger than their desire (if any) to be epistemically rational” (Huemer, 2015).
20. See also her (2005) and (2006).
21. Hieronymi attributes this view to Kolodny in a footnote. To my knowledge, he has not defended it in print. His answer is very similar to something Sher (2021: 24) has defended in print, however.
22. Some, such as Sher (2021: 24) give them separate cost-benefit analyses. The difficulties for such a position should be clear in light of the above.
23. Thanks to an anonymous referee for pointing out the need to address this point.
24. Thanks to an anonymous referee for raising this question.

References

- Adler J and Hicks M (2013) Non-evidential reasons to believe. In: Chan T (eds) *The Aim of Belief*. Oxford: Oxford University Press, pp. 140–166.
- Ariely D, Loewenstein G, and Prelec D (2003) Coherent arbitrariness: stable demand curves without stable preferences. *The Quarterly Journal of Economics* 118(1): 73–105.
- Berker S (2013) Epistemic teleology and the separateness of propositions. *Philosophical Review* 122: 337–393.
- Boghossian P (2003) The normativity of content. *Philosophical Issues* 13: 31–45.
- Brewer T (2009) *The Retrieval of Ethics*. Oxford: Oxford University Press.
- Bykvist K and Hattiangadi A (2007) Does thought imply ought. *Analysis* 67: 277–285.
- Bykvist K and Hattiangadi A (2013) Belief, truth and blindspots. In: Chan T (eds) *The Aim of Belief*. Oxford: Oxford University Press, pp. 100–122.
- Caplan B (2001) Rational ignorance. *Kyklos* 54: 3–26.
- Caplan B (2006) *The Myth of the Rational Voter: Why Democracies Choose Bad Policies*. Princeton: Princeton University Press.
- Cohen GL (2003) Party over policy: the dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology* 85: 808–822.
- Coleman JS (1964) *Introduction to Mathematical Sociology*. New York: Free Press of Glencoe.
- DeRose K (2002) Assertion, knowledge and context. *Philosophical Review* 111: 267–203.
- Downs A (1957) *An Economic Theory of Democracy*. New York: Harper and Row.
- Easwaran K (2013) Expected accuracy supports conditionalization—and conglomerability and reflection. *Philosophy of Science* 80: 119–142.
- Easwaran K and Fitelson B (2015) Accuracy, coherence and evidence. In: Gendler T and Hawthorne J (eds) *Oxford Studies in Epistemology*, Vol. 5. Oxford: Oxford University Press, pp. 61–96.
- Engel P (2005) Truth and the aim of belief. In: Gillies D (eds) *Laws and Models in Science*. London: King’s College Publications.
- Engel P (2007) Belief and normativity. *Disputatio* 2(23): 197–202.

- Engel P (2013) In defense of normativism about the aim of belief. In: Chan T (eds) *The Aim of Belief*. Oxford: Oxford University Press, pp. 32–63.
- Evans G (1982) *The Varieties of Reference*. Oxford: Clarendon Press.
- Feldman R (2000) The ethics of belief. *Philosophy and Phenomenological Research* 60: 667–696.
- Gibbons J (2013) *The Norm of Belief*. Oxford: Oxford University Press.
- Gluer K and Wikforss Å (2009) Against content normativity. *Mind; A Quarterly Review of Psychology and Philosophy* 118: 31–70.
- Greene JD (2013) *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: Penguin Press.
- Hannon M (2022) Are knowledgeable voters better voters? *Politics, Philosophy & Economics* 21(1): 29–54.
- Hieronymi P (2005) The wrong kind of reason. *The Journal of Philosophy* 102(9): 435–457.
- Hieronymi P (2006) Controlling attitudes. *Pacific Philosophical Quarterly* 87(1): 45–74.
- Hieronymi P (2009) Two kinds of agency. In: O’Brien L and Soteriou M (eds) *Mental Actions*. Oxford: Oxford University Press, pp. 138–162.
- Homans GC (1958) Social behavior as exchange. *American Journal of Sociology* 63(6): 597–606.
- Horwich P (2013) Belief-truth norms. In: Chan T (ed) *The Aim of belief*. Oxford, UK: Oxford University Press, pp. 17–31.
- Huemer M (2015) Why we are irrational about politics. In: Anomaly J, Brennan G, Munger M, et al (eds) *Philosophy, Politics and Economics: An Anthology*. Oxford, UK: Oxford University Press, 456–467.
- Humberstone L (1992) Direction of fit. *Mind* 101(401): 59–83.
- Jost JT, Hennes EP, and Lavine H (2013) Hot political cognition: its self-, group-, and system-serving purposes. In: Carlson DE (eds) *Oxford Handbook of Social Cognition*. Oxford: Oxford University Press, pp. 851–875.
- Joyce J (1998) A non-pragmatic vindication of probabilism. *Philosophy of Science* 65: 575–603.
- Kahan D (2013) Ideology, motivated reasoning and cognitive reflection. *Judgment and Decision Making* 8(4): 407–424.
- Kahan D *Emerging Trends in the Social and Behavioral Sciences*. <http://onlinelibrary.wiley.com/doi/10.1002/9781118900772.etrds0418/abstract>
- Kahan D *Emerging Trends in the Social and Behavioral Sciences*. <http://onlinelibrary.wiley.com/doi/10.1002/9781118900772.etrds0417/pdf>
- Kahneman D and Tversky A (1982) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge, MA: Cambridge University Press.
- Kavka G (1983) The toxin puzzle. *Analysis* 43: 33–36.
- Landemore H (2012) Democratic reason: the mechanisms of collective intelligence in politics. In: Landemore H and Elster J (eds) *Collective Wisdom: Principles and Mechanisms*. New York: Cambridge University Press, pp. 251–290.
- Leitgeb H and Pettigrew R (2010) An objective justification of Bayesianism I: measuring inaccuracy. *Philosophy of Science* 77: 201–235.
- Mackie G (2012) Rational ignorance and beyond. In: Landemore H and Elster J (eds) *Collective Wisdom: Principles and Mechanisms*. New York: Cambridge University Press, pp. 290–319.
- Madrian B and Shea D (2001) The power of suggestion: inertia in 401(k) participation and savings behavior. *The Quarterly Journal of Economics* 116(4): 1149–1187.

- Millar A (2004) *Understanding People: Normativity and Rationalizing Explanation*. Oxford: Oxford University Press.
- Moran R (1988) Making up your mind: self-interpretation and self-constitution. *Ratio* NS 1: 135–151.
- Moran R (2001) *Authority & Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press.
- Olson M (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge: Harvard University Press.
- Owens D (2003) Does belief have an aim?. *Philosophical Studies* 115(3): 283–305.
- Papineau D (2013) There are no norms of belief. In: Chan T (ed) *The Aim of belief*. Oxford, UK: Oxford University Press, 64–79.
- Plato (2004) *Republic*. Trans. Reeve CDC. Indianapolis, IN: Hackett Publishing Company.
- Reisner A (2007) Evidentialism and the numbers game. *Theoria* 73: 304–316.
- Reisner A (2009) The possibility of pragmatic reasons for belief and the wrong kind of reasons problem. *Philosophical Studies* 145: 257–272.
- Reisner A (2013) Leaps of knowledge. In: Chan T (eds) *The Aim of Belief*. Oxford: Oxford University Press, 167–183.
- Rinnard S (2017) No exception for belief. *Philosophy and Phenomenological Research* 94(1): 121–143.
- Rinnard S (2019) Equal treatment for belief. *Philosophical Studies* 176(7): 1923–1950.
- Schelling T (1960) *The Strategy of Conflict*. Cambridge: Harvard University Press.
- Schumpeter J (1950) *Capitalism, Socialism and Democracy*. New York: Harper and Brothers.
- Shah N (2003) How truth governs belief. *Philosophical Review* 112: 447–482.
- Shah N (2004) A new argument for evidentialism. *The Philosophical Quarterly* 225(56): 481–498.
- Shah N and Velleman D (2005) Doxastic deliberation. *Philosophical Review* 114: 494–534.
- Sher G (2021) *A Wild West of the Mind*. Oxford, UK: Oxford University Press.
- Sherman DK and Cohen GL (2006) The psychology of self-defense: self-affirmation theory. *Advances in Experimental Social Psychology* 38: 183–242.
- Somin I (2013) *Democracy and Political Ignorance: Why Smaller Government is Better*. Redwood City, CA: Stanford University Press.
- Steglich-Peterson A (2006) No norm needed: on the aim of belief. *The Philosophical Quarterly* 56: 499–516.
- Steglich-Peterson A (2009) Weighing the aim of belief. *Philosophical Studies* 145: 395–405.
- Vahid H (2006) Aiming at truth: doxastic vs. epistemic goals. *Philosophical Studies* 131: 301–335.
- Velleman JD (2000) *The Possibility of Practical Reason*. Oxford: Oxford University Press, pp. 224–282.
- Wedgwood R (2002) The aim of belief. *Philosophical Perspectives* 16: 267–297.
- Wedgwood R (2013) The Right thing to Believe. In: Chan T (ed) *The Aim of Belief*. Oxford: Oxford University Press, pp. 123–139.
- Whiting D (2012) Does belief aim (only) at the truth. *Pacific Philosophical Quarterly* 93: 279–300.
- Whiting D (2013) Nothing but the truth: on the norms and aims of belief. In: Chan T (eds) *The Aim of Belief*. Oxford: Oxford University Press, pp. 183–203.
- Williams B (1973) Deciding to believe. In *Problems of the Self*. Cambridge: Cambridge University Press, pp. 136–151.
- Zalabardo J (2010) Why believe the truth?: Shah and Velleman on the aim of belief. *Philosophical Explorations* 13: 1–21.

Author Biography

Spencer Paulson is a PhD candidate at Northwestern University in the Philosophy Department. He primarily works on Epistemology and the Philosophy of Mind.