THE JOURNAL OF PHILOSOPHY

VOLUME CXVI, NUMBER 5
MAY 2019

***Published by* The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY
VOLUME CXVI, NO. 5, MAY 2019

THE TRUTH PROBLEM FOR PERMISSIVISM*

Epistemologists often assume that if rationality is worth pursuing, it must bear some sort of connection to the truth. What exactly this connection amounts to is mysterious, but the thought that there must *be* such a connection seems to limit our theory of rationality in various ways. For instance, a classic objection to coherentism is that the view seemingly has no safeguards against rational believers who get things very wrong—so, one might think, the demand for a truth-connection favors externalist views over internalist views. In formal epistemology, various understandings of the truth-connection have been used to argue for formal norms such as probabilism and conditionalization. This paper will examine the truth problem as it relates to *permissivism*.[1] If rationality is a guide to the truth, can it also allow some leeway in how we should respond to our evidence?

In the first half of the paper I will look at a particular strategy for connecting *permissive* rationality and the truth, developed in a recent

*Many thanks to David Christensen, Elizabeth Harman, Hilary Kornblith, Matt Mandelkern, Miriam Schoenfield, Henry Swift, and audiences at the University of Toronto, Fordham University, SWIP-Analytic, and UT Austin for valuable discussion and suggestions. I would also like to thank two anonymous referees for their helpful and extensive comments.

[1] The relationship between permissivism and the truth-connection has been discussed in a few recent papers. See Sophie Horowitz, "Immoderately Rational," *Philosophical Studies*, CLXVII, 41 (January 2014): 41–56; Benjamin Anders Levinstein, "Permissive Rationality and Sensitivity," *Philosophy and Phenomenological Research*, XCIV, 2 (March 2017): 342–70; Sinan Dogramaci and Sophie Horowitz, "An Argument for Uniqueness About Evidential Support," *Philosophical Issues*, XXVI, 1 (October 2016): 130–47; Daniel Greco and Brian Hedden, "Uniqueness and Metaepistemology," this JOURNAL, CXIII, 8 (August 2016): 365–95; and Miriam Schoenfield, "Permissivism and the Value of Rationality: A Challenge to the Uniqueness Thesis," *Philosophy and Phenomenological Research* (2018). Roger White discusses some related issues; see Roger White, "Epistemic Permissiveness," *Philosophical Perspectives*, XIX, 1 (December 2005): 445–59. This paper will most closely engage with the arguments in Schoenfield, "Permissivism and the Value of Rationality," *op. cit.*, and Horowitz, "Immoderately Rational," *op. cit.*

0022-362X/19/1605/237–262

paper by Miriam Schoenfield. This strategy says (roughly) that there are limits on what we regard as rational, and therefore there are also limits on the extent to which we should regard rationality as leading to the truth. However, Schoenfield argues, there is a sense in which permissivism can deliver a truth-connection. I will argue that this limited truth-connection is unsatisfying, and the version of permissivism that supports it faces serious challenges; so, for mainstream permissivism, the truth problem is still unsolved. In the second half of the paper I will look at a strategy available to *impermissivists*, according to which rationality bears a quite strong connection to truth. I will argue that this second strategy is successful.

I. PERMISSIVISM AND TRUTH

Permissivism, as I will understand it, is roughly the view that there can be rational disagreements on the basis of a single body of evidence. According to *impermissivism* (also called "Uniqueness"), this is not possible: there can be no reasonable disagreements without differences in evidence. I will define impermissivism more precisely as follows:

> IMPERMISSIVISM: For any body of evidence E, and proposition P, there is at most one doxastic attitude toward P that is consistent with being ideally rational and having E as one's total evidence.

Permissivism, then, is the view that impermissivism is false: sometimes, multiple doxastic attitudes toward P are consistent with being ideally rational and having E as one's evidence.[2]

Of course, no plausible theory of rationality can *guarantee* a truth-connection. Evidence is sometimes misleading, and rational beliefs are sometimes false. The arguments I will focus on each set their sights a bit lower, claiming that we should *defeasibly expect* rationality (according to some given theory) to lead us to the truth. In this section I will introduce one attempt to show how, given a certain permissive notion of rationality, we can expect rationality to give us accurate beliefs.

The permissivist argument I will examine starts with a thought commonly taken to motivate permissivism: that *rationality only takes us so far*. In some circumstances, rationality recommends one belief state

[2] Various participants in the literature understand "permissivism" and "impermissivism" in subtly different ways. The definition I will use here is based on Schoenfield's definition (see Schoenfield, "Permissivism and the Value of Rationality," *op. cit.*) with one difference. Where Schoenfield defines impermissivism as always requiring a unique doxastic attitude toward any proposition, my definition leaves it open that for some propositions, no doxastic attitude is ideally rational. For discussion of the differences between various forms of permissivism and impermissivism, see Matthew Kopec and Michael Titelbaum, "The Uniqueness Thesis," *Philosophy Compass*, XI, 4 (April 2016): 189–200.

or another, but in other circumstances it does not. If rationality only takes us so far, it can only take us so far *toward* anything, including the truth. So, someone might endorse the following rough picture of how rationality is connected to truth:

> PERMISSIVIST ARGUMENT (VERY ROUGH PASS): Sometimes, rationality tells us what to believe. In those circumstances we should expect that it is pointing us to the truth. But other circumstances are beyond the scope of the rational rules, and so, rationality does not tell us what to believe. We should therefore endorse the following sort of connection between rationality and truth: Insofar as it points anywhere, rationality points to the truth.

To defend this permissivist argument, one must give support for the following two claims:

> CLAIM 1: When rationality tells us what to believe, usually what it tells us to believe is true.
>
> CLAIM 2: Rationality often does not tell us what to believe.

The first claim is about the truth-connection itself; the second is one way of stating permissivism. (I am using "belief" here to include credences as well. One could reformulate Claims 1 and 2 to be about credences and accuracy rather than belief and truth.) To see how one might support both claims together, let us turn to the picture developed recently by Schoenfield. In the next section I will evaluate this general permissivist strategy, as well as Schoenfield's more specific version of it.

Schoenfield's defense of these two claims can be broken down into three pieces. Together I will call them the "Endorsement Argument." (What follows is my reconstruction of her argument.) The first piece is a purported data point, which is that we (meaning, literally, you and I) endorse certain belief-forming rules.[3] These might include logical rules like modus ponens, or more substantive rules such as "trust your perception," "reason inductively rather than counterinductively," and so on. (It does not matter what the rules are, as long as there are some.) To "endorse" these rules, in the relevant sense, is to regard them as truth-conducive: these rules are what we prefer from an accuracy-seeking perspective. (Throughout, I will use "accuracy" to denote closeness to the truth, and sometimes switch between talk of

[3] Schoenfield calls these "cognitive properties." She specifies that they must be "specifiable in purely descriptive language" and that they must "[supervene] on the agent's non-factive mental states" (Schoenfield, "Permissivism and the Value of Rationality," *op. cit.*, p. 3). The point of these requirements is to make sure that rules like "believe rationally" and "believe truly" do not count. Cognitive properties include specific attitudes as well as belief-forming rules or methods.

accuracy and talk of truth.) Note that endorsement is a feature of our preferences; it need not be conscious or explicit.[4]

The second piece is another purported data point: that the rules we endorse are not exhaustive. Instead, they sometimes do not yield any answer about what to believe. For example, Schoenfield asks: what is your credence that it will rain in Honolulu next New Year's Day? Plausibly, she argues, you *have* no precise credence here, and there is also no credence that you think would be particularly accurate or truth-conducive. This means that the rain-in-Honolulu case is permissive: many possible attitudes toward rain-in-Honolulu are compatible with following the rules you endorse.

The last piece of the Endorsement Argument associates *endorsing* with *regarding as rational*. On this view, what you judge to be rational is just what you judge to be compatible with all the belief states and belief-forming rules that you endorse. Seeing endorsement and rationality judgments as closely connected fits nicely with certain metaepistemological views, such as expressivism. Although Schoenfield does not explicitly defend taking this step, it is strongly suggested at the end of her paper. (More precisely, she argues that we would not want to make rationality judgments that go beyond endorsement—that is, we would not want to judge something to be uniquely rational without also endorsing it.) Since I am primarily interested in the general permissive strategy that holds Claims 1 and 2, from above, I will focus on what we might say if we associate endorsement with regarding as rational.[5]

If we put together the three pieces above, we can support Claims 1 and 2. It is easy to see how we get Claim 1. If the epistemic rules and belief states that we regard as rational *just are* those we regard as truth-conducive—or compatible with the epistemic rules and belief states that we regard as truth-conducive—then of course we should hold that rationality tends to lead to the truth. Claim 2 is straightforward as well: rationality does not always tell us what to believe simply because the rules we endorse do not extend to all cases. For example, these rules deliver no judgment about how likely it is to rain in Honolulu on

[4] "We'll say that an agent *endorses a set of cognitive properties*, C, if she prefers, *when her only goal is accuracy*, any cognitive system that instantiates all of the properties in C, to a cognitive system that lacks some of these properties." (*Ibid.*, p. 3.)

[5] Note that this is not the view that whatever we endorse is rational. Schoenfield draws an analogy with expressivism here: for an expressivist, having some moral judgment entails having some other attitude, such as approval or disapproval. However, Schoenfield points out, "[T]he expressivist is not a relativist" (*ibid.*, p. 4). Schoenfield's aim is therefore to bring out some features of what we endorse, and what we judge to be rational, rather than to draw direct conclusions about rationality itself.

such-and-such future date. This appears to give us what we were after: a connection between permissive rationality and truth.

II. TWO OBJECTIONS AND A QUESTION FOR THE PERMISSIVIST ARGUMENT

The strategy outlined above seems promising. However, as we will see, it is not the good news for permissivism that it might appear to be. The connection between rationality and truth is relatively weak—meaning that while rationality and truth are conceptually linked, on this view, rational agents should not think that rationality is always better than irrationality for the purposes of getting to the truth. More worryingly, the truth-connection offered by this argument is only available to a narrow class of permissive views, which are independently unattractive.

These two limitations, which I will discuss in sections II.1 and II.2, stem from a common source: the fact that *holding* some doxastic attitude necessarily involves regarding that attitude as more accurate than the alternatives. (Believing that P involves regarding P as true; it would therefore be incoherent for someone to believe P but also regard the belief that $\sim P$ as more accurate than the belief that P.) This thesis, called "Immodesty," is widely held by epistemologists (including Schoenfield, in situations in which our credences are precise[6]); formal epistemologists often take Immodesty as a constraint on acceptable accounts of accuracy. Immodesty can help explain why, for example, a rational agent with .6 credence in P will prefer to stay at her .6 credence (absent new evidence) rather than switch to a different credence. Extending this thought beyond an agent's current credences to her epistemic rules or plans, rational agents will also expect their responses to various bodies of evidence (assuming those conform to their plans) to maximize expected accuracy, given the evidence.[7]

[6] *Ibid.*, pp. 4–5, 9, and 11 (note 16). Though she does not use the term "Immodesty," Schoenfield discusses, with approval, the idea that someone who adopts a certain doxastic attitude should regard that attitude as optimal with regard to accuracy. In an earlier paper, Schoenfield appeals to Immodesty more directly. (Miriam Schoenfield, "Permission to Believe: Why Permissivism Is True and What It Tells Us About Irrelevant Influences on Belief," *Noûs*, XLVIII, 2 (June 2014): 193–218.) In cases of imprecise or "mushy" credences, Schoenfield has argued that Immodesty does not hold. (Miriam Schoenfield, "The Accuracy and Rationality of Imprecise Credences," *Noûs*, LI, 4 (December 2017): 667–85.) My focus here will be in applying Schoenfield's more recent proposal (in Schoenfield, "Permissivism and the Value of Rationality," *op. cit.*) to possible permissive situations in which we do have precise credences. I discuss this further at the end of this subsection.

[7] I will not defend Immodesty here for reasons of space, but for further discussion of Immodesty in this context, see Horowitz, "Immoderately Rational," *op. cit.* Immodesty is typically accepted in the literature on epistemic utility theory. See, for example, James M. Joyce, "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief," in Franz Huber and Christoph Schmidt-Petri, eds., *Degrees of Belief* (Dordrecht, the Netherlands: Springer, 2009), pp. 263–97; Hilary Greaves and David Wallace, "Justifying

Because of Immodesty, notice that a rational agent will take just one attitude toward a given proposition (her own attitude) to be the best in terms of accuracy. And this means she will only "endorse" one attitude, in Schoenfield's sense. Insofar as she regards other attitudes as more or less accurate, she will give better marks to attitudes that are closer to her own. (For example, if her credence in P is .6, then she will regard a credence of .61 as more accurate than a credence of .7.) These implications of Immodesty will make trouble for the Endorsement Argument, both in articulating a strong connection between rationality and truth, and in accommodating "acknowledged permissive cases."

Before proceeding, let me make a note about my focus here. I am interested in cases where agents have *precise* credences in response to a permissive body of evidence. In taking this focus I am departing somewhat from Schoenfield's discussion, which for the most part focuses on cases in which we have either imprecise credences or no credences at all in response to a permissive body of evidence. However, Schoenfield does not rule out the possibility that we will sometimes hold precise credences in permissive cases.[8] More importantly, other permissivist epistemologists typically focus on cases in which we have precise credences. For instance, Kelly argues that permissivism becomes *more* plausible when one thinks of belief in a fine-grained way.[9] And others routinely assume, either in examples or in offhand comments, that it can be permissible to have precise credences in these cases.[10] For those reasons, I take it that permissivists in general should be interested in what Schoenfield's argument has to say about such cases; if this argument turns out to lead to problems, this will give mainstream permissivists reason not to take it on board.[11]

Throughout sections II.1 and II.2 I will assume, as Schoenfield does, that this version of permissivism is compatible with probabilism as a rational constraint. I will examine this assumption more closely in section II.3. (So II.1 and II.2 will contain objections; II.3 will raise a question.)

II.1. The Connection between Permissive Rationality and Truth. Schoenfield writes that "it's no mystery" why one would want to be rational in her permissive sense: being rational just amounts to believing in accordance with the rules that one takes to be truth-conducive.[12] But as I will argue in this section, there is still a bit of mystery left: the connection between rationality and truth (on Schoenfield's permissive view) is not very strong. Rationality is a way to get to the truth, but not a very good way. I will bring this out by showing that if we are rational, it will not always be the case that we expect particular rational credences supported by our evidence to be more accurate than particular irrational credences.[13]

As an example, consider the perspective of a rational agent, Ruth, who has precise opinions regarding two propositions P and Q. Ruth's evidence is *impermissive* regarding P (perhaps it includes information about objective chances) and *permissive* regarding Q. Suppose Ruth knows that her credences in P and Q are rational, and also knows that some other credences in Q are rational besides her own. (In section II.2 I will question how she can hold this latter attitude, given Schoenfield's picture, but let us assume she can for now.) Finally,

Conditionalization: Conditionalization Maximizes Expected Epistemic Utility," *Mind*, CXV, 459 (July 2006): 607–32; and Richard Pettigrew, *Accuracy and the Laws of Credence* (Oxford: Oxford University Press, 2016). The argument I am presently considering from Schoenfield also accepts Immodesty as applied to precise credences. The distinction between beliefs and plans or rules is not always so clear—for example, one's *plan* to adopt the belief that all emeralds are green, after seeing n green emeralds, should be closely related to one's *belief* about how much variability there is in emerald color—so it makes sense for Immodesty to apply to both.

[8] See especially Schoenfield, "Permissivism and the Value of Rationality," *op. cit.*, p. 11, fn 16. There, she discusses which rationality judgments one might make if one has adopted a precise credence in a permissive case.

[9] Thomas Kelly, "Evidence Can Be Permissive," in Matthias Steup, John Turri, and Ernest Sosa, eds., *Contemporary Debates in Epistemology*, 2nd ed. (Malden, MA: Wiley-Blackwell, 2014), pp. 298–312.

[10] To take just a few examples, see the fine-grained "Reasoning Room" case from Michael Titelbaum and Matthew Kopec, "When Rational Reasoners Reason Differently," in Magdalena Balcerak-Jackson and Brendan Balcerak-Jackson, eds., *Reasoning: Essays on Theoretical and Practical Thinking* (Oxford: Oxford University Press, forthcoming); "Case 3" and "Case 4" from Thomas Kelly, "Peer Disagreement and Higher-Order Evidence," in Richard Feldman and Ted A. Warfield, eds., *Disagreement* (Oxford: Oxford University Press, 2010), pp. 111–74; and discussion of Kelly's example and others in Nathan Ballantyne and E. J. Coffman, "Conciliationism and Uniqueness," *Australasian Journal of Philosophy*, XC, 4 (2012): 657–70.

[11] It is also important to note that I do not intend my arguments here to apply to cases in which we have imprecise credences or no credences at all. That is because, as Schoenfield has argued elsewhere (see Schoenfield, "The Accuracy and Rationality of Imprecise Credences," *op. cit.*), imprecise credences do not support Immodesty. Schoenfield writes, "[I]f imprecise credal states are made rational by a certain kind of evidence, this is not a fact that can be explained by our interest in having doxastic states that accurately represent the way the world is" (*ibid.*, p. 680). She also suggests that we might want to deny that imprecise credences can ever be rational (or even rationally permissible). So while a permissivist could sidestep the arguments here by denying that we are ever permitted to have *precise* credences, this strategy is not without its challenges. Thanks to an anonymous referee for pressing me to clarify this issue.

[12] Schoenfield, "Permissivism and the Value of Rationality," *op. cit.*, p. 7. At this point in the paper Schoenfield is discussing a stipulated property, "rotionality," but she later suggests that rotionality and rationality are equivalent.

[13] The argument in this section is based on a similar argument from Horowitz, "Immoderately Rational," *op. cit.*, p. 50. That argument is directed toward a different sort of permissive view, but the same kind of objection seems to apply here.

suppose Ruth is *immodest*. She takes her own attitudes to have the best shot at accuracy; she regards alternative attitudes to be less accurate the farther they are from her own.

Now let us suppose Ruth's two friends, Adam and Roma, share her evidence regarding *P* and *Q*. Roma shares Ruth's credence in *P* but has a significantly higher, though also rational, credence in *Q*. Adam shares Ruth's rational credence in *Q* but has a very slightly different credence in *P*. If Ruth compares her credences to her friends' credences, she will come to the following conclusion: Roma is more *rational* than Adam, but Adam is more *accurate* than Roma.[14] If you are in Ruth's position, it is indeed a mystery why you would prefer Roma's credences over Adam's, for any accuracy-related reasons.

If one endorses a set of belief-forming rules that allows for both impermissive and permissive cases, then situations like Ruth's will be inevitable. Rational agents will sometimes regard *irrational* responses to a given body of evidence as *more accurate* than rational responses.[15]

What this means is that while the Endorsement Argument secures a connection between rationality and truth, for permissivism, it turns out that rationality is still not a very good way to get to the truth. (It is the best we can do, according to this argument, but the best we can do is not great.) So while what is good about rationality is not a mystery, one might still wonder what is especially good about rationality.

[14] For a more concrete example, we can assign some numbers. (Adam's credence in *P* is the only irrational credence of the six listed.)

	Credence in *P*	Credence in *Q*
Ruth	.6	.1
Roma	.6	.5
Adam	.61	.1

It is easy to set up situations like this. Ruth will regard Adam's credences as more expectedly accurate than Roma's if the difference in expected accuracy (from Ruth's point of view) between Ruth and Roma, regarding *Q*, is greater than the difference in expected accuracy between Ruth and Adam, regarding *P*. In this example, assuming that all three agents have probabilistic credences, then, using the Brier score: from Ruth's point of view, the expected inaccuracy of Adam's credences is $(.2841+.09)/2 = .18705$ and the expected inaccuracy of Roma's credences is $(.24+.25)/2 = .245$.

[15] In fact, such situations might come about even if one's permissive rules did not allow for any impermissive cases. Someone who only endorsed probabilism could end up in a situation like Ruth's, regarding some nonprobabilistic (and hence irrational) credences as more accurate than some probabilistic (and hence rational) credences. I thank an anonymous referee for raising this point. I focus on this sort of case because I am unsure how to best understand someone who endorses only probabilism on this kind of view. See section II.3 for further discussion of how we might defend probabilism (or other consistency requirements) in the context of the Endorsement Argument.

(Compare: suppose I say to you, "It's no mystery why someone would want to eat a peanut butter sandwich. Peanut butter sandwiches are healthy and taste good!" You might reply that it is still mysterious why someone would *choose* to eat a peanut butter sandwich, or make a peanut butter sandwich for a friend, if there are other available options. This is because many foods are both healthier and more delicious than a peanut butter sandwich. Similarly, if Ruth cares about accuracy, she will be able to say something good about Roma's credences. But Adam's will look even better.) This is of course not a fatal flaw by itself—for all we know so far, perhaps this permissive strategy is the best we can do. However, it does give us reason to look for something better.

II.2. Acknowledged Permissive Cases. A more interesting and pressing problem for this version of permissivism has to do with the possibility of *knowing* that one is in a permissive situation.[16] (In the previous subsection I assumed that the Endorsement Argument was compatible with the existence of acknowledged permissive cases; now it is time to question that.) One of the main benefits of permissivism is its purported ability to explain situations in which people can "agree to disagree"—about politics, religion, jury verdicts, and so forth—yet still respect one another's epistemic credentials.[17] If permissivism is to reap this benefit, it must be able to explain how rational agents can sometimes judge that their own belief about some matter is just one of several permissible options. In this section I will show why permissivists will have trouble accommodating *both* acknowledged permissive cases and Schoenfield's Endorsement Argument.

[16] Though for an exception, see Stewart Cohen, "A Defense of the (Almost) Equal Weight View," in David Christensen and Jennifer Lackey, eds., *The Epistemology of Disagreement: New Essays* (Oxford: Oxford University Press, 2013), pp. 98–117. Schoenfield (Schoenfield, "Permissivism and the Value of Rationality," *op. cit.*) says that she is officially "neutral" on this issue but does suggest a way in which rational agents could regard their own positions as permissive. See Ballantyne and Coffman, "Conciliationism and Uniqueness," *op. cit.*; and Titelbaum and Kopec, "When Rational Reasoners Reason Differently," *op. cit.*, for further discussion of acknowledged permissive cases.

[17] For example, consider this quotation from Rosen (Gideon Rosen, "Nominalism, Naturalism, Epistemic Relativism," *Philosophical Perspectives*, XV, 1 (2001): 69–91, at pp. 71–72): "It should be obvious that reasonable people can disagree, even when confronted with a single body of evidence. . . .Paleontologists disagree about what killed the dinosaurs. And while it is possible that most of the parties to this dispute are irrational, this need not be the case. To the contrary, it would appear to be a fact of epistemic life that a careful review of the evidence does not guarantee consensus even among thoughtful and otherwise rational investigators." Presumably these paleontologists have systematic differences in their scientific commitments, rather than one-off random disagreements. Schoenfield's earlier work also takes religious disagreement as a paradigmatic example of permissivism. (See Schoenfield, "Permission to Believe," *op. cit.*)

I will look at two ways in which a permissivist might make sense of acknowledged permissive cases. The first way is available to many different permissivist views but cannot appeal to the Endorsement Argument. The second is compatible with the Endorsement Argument but leads us to an implausible view about acknowledged permissive cases.

For both options, let us focus on a specific (purported) permissive case: again, the proposition that it will rain in Honolulu next New Year's Day. Call that "*H*." Suppose a rational agent, Petra, has evidence *E*, has adopted some rational attitude toward *H* (one of the many rational attitudes), and is now considering whether her evidence is permissive regarding *H*. If Petra's current situation is an *acknowledged permissive case*, then she should be able to reach the conclusion that her own belief is just one of many permissible alternative beliefs. (For the arguments in this section it does not matter whether Petra can know what those other beliefs are.)

The first option is one I will call "Personal Rules" permissivism. In the literature this is sometimes referred to as the view that rationality is "interpersonally permissive, but intrapersonally impermissive." For each person, on this type of view, rationality mandates a specific response to any given body of evidence. But this response varies from person to person. Personal Rules permissivists may allow for a large number of impermissive situations, where all rational rules agree on what the evidence supports. But they also allow for situations in which different agents' personal rules disagree; those are the permissive cases, where rationality by itself does not dictate how one should believe.

Examples of Personal Rules permissivism include subjective Bayesianism (according to which each agent must update by conditionalization, starting from her own initial credences) as well as some informal permissive views, according to which rationality allows a range of ways to respond to one's evidence determined by priorities regarding believing truth and avoiding error, "power" versus "reliability," and so forth.[18] It also fits well with the view that we have our own "epistemic standards," as developed in an earlier defense of permissivism by Schoenfield.[19]

Personal Rules permissivists can explain acknowledged permissive cases precisely because they see a distinction between an agent's judging a belief to be *rational* in response to some evidence and judging it to be *accurate* given that evidence. Let us return to Petra. If Petra's belief about *H* is rational, according to Personal Rules permissivism, that is

[18] See, for example, Kelly, "Evidence Can Be Permissive," *op. cit.*; and Hartry Field, "Apriority as an Evaluative Notion," in Paul Boghossian and Christopher Peacocke, eds., *New Essays on the A Priori* (Oxford: Oxford University Press, 2000), pp. 117–49.

[19] Schoenfield, "Permission to Believe," *op. cit.*

because it accords with Petra's personal rule. And because of Immodesty, Petra should take her attitude in *H* to be the most *accurate* one, given her evidence. But since this is a permissive case, it is one where Petra's personal rule goes *beyond* the rational requirements. If Petra can recognize this fact, then she should be able to acknowledge that she is in a permissive case: her own attitude toward *H* is permissible, but so are others. This strategy can explain how someone like Petra can recognize herself as being in a permissive case (she can see that many options are rational) and yet maintain her own belief in response to her evidence (she sees others as less accurate). In fact, Schoenfield's earlier work uses this very argument to explain acknowledged permissive cases.

However, this argument is incompatible with the Endorsement Argument—precisely because the Endorsement Argument requires that endorsing-as-accurate entails endorsing-as-uniquely-rational. If an agent has her own personal rules, then those are the rules she will "endorse" in Schoenfield's sense; therefore, in any situation that is governed by an agent's personal rule, the agent will take that situation to be *impermissive*.[20]

This means that the Endorsement Argument is *not* available to views that are interpersonally permissive but intrapersonally impermissive. Since interpersonal permissivism and intrapersonal impermissivism are a popular combination, this limitation is significant.[21]

Let us turn to a different possibility for accommodating acknowledged permissivism. This one *is* compatible with the Endorsement Argument—it is Schoenfield's own suggestion. Remember that Schoenfield's preferred version of permissivism is one on which, in permissive cases, the rules we endorse do not determine what we should believe. So if Petra happens to adopt some attitude about *H*, it will come about through some non-rule-governed process: a bump on the head, random guessing, or something like that. In other words, Petra's attitude about *H* is *not* recommended by any of the epistemic rules that she endorses. However, once again because of Immodesty, once Petra has a precise opinion about *H*, she will *endorse* that opinion as accurate (and hence regard it as rational). So how can Petra recognize that her case is permissive?

To allow for acknowledged permissivism in situations like Petra's, Schoenfield suggests a slight modification to her account of

[20] Schoenfield expressly does not defend this view in her "Permissivism and the Value of Rationality."

[21] See, for instance, Christopher J. G. Meacham, "Impermissive Bayesianism," *Erkenntnis*, LXXIX, Supplement 6 (June 2014): 1185–217; and Kelly, "Evidence Can Be Permissive," *op. cit.*, for discussion of the difference between interpersonal and intrapersonal permissivism. Both authors defend permissivism, and both suggest that interpersonal permissivism is more plausible than intrapersonal permissivism.

endorsement. She writes that if Petra *sets aside* her opinion about *H*, she will be able to recognize that there are *many* belief states, including her opinion about *H*, that are not ruled out by any of the rules she endorses.[22] So she will regard her situation as permissive. Let us call this the "Setting-Aside Strategy."

The Setting-Aside Strategy gives us a way to allow for acknowledged permissive cases *and* accept the Endorsement Argument. But permissivists should be hesitant to accept the resulting view. That is because, while this suggestion delivers acknowledged permissivism in one-off cases like the rain-in-Honolulu example, it does not work in many of the paradigmatic (purported) permissive cases that many epistemologists want to recognize. As I mentioned above, permissivists often argue that many permissive cases involve entire sets or systems of beliefs: religious or political worldviews, scientific traditions, or methods of weighting various broad types of evidence. Permissivism is (according to many epistemologists) supposed to legitimize a certain kind of epistemic tolerance in the face of such systematic disagreement. Thus, it is bad news if our view says that such situations are permissive but that we cannot acknowledge them as such.

To see why broader, acknowledged permissive cases are impossible, using the Setting-Aside Strategy, suppose that Petra has a comprehensive set of religious beliefs, which comprise one of many rationally permissible religious belief systems. Now suppose Petra picks out just one of her religious beliefs, *R*, and asks herself whether her own epistemic situation is permissive regarding *R*. Since she already has an opinion, she needs to set it aside to see whether she endorses it independently of holding it; suppose she does set it aside. What will happen? Well, if her religious beliefs are at all systematic, it seems that she will be able to recover the set-aside belief, *R*, by using the rest of her beliefs—just as she would be able to recover beliefs formed through the rational rules. The *rest* of Petra's beliefs, and in particular the religious beliefs that she has *not* set aside, will favor her set-aside belief over the alternatives. So she will regard her evidence about *R* as *impermissive*.

To take a more concrete example, suppose Petra's religious beliefs are roughly those recommended by Catholicism, and *R* is the proposition that the Holy Communion is literally the body and blood of Christ. The result that most permissivists want is for Petra to be able to say something like this: "I believe that Holy Communion is literally the body and blood of Christ; however, I recognize that others with my very same evidence may rationally believe that it is merely an expression of faith, or a symbolic

[22] See Schoenfield, "Permissivism and the Value of Rationality," *op. cit.*, p. 16, note 11.

reenactment of the Last Supper." But if the Setting-Aside Strategy is how Petra is to check whether this case is permissive, we will not get the result permissivists want. The Catholic view of Holy Communion is of a piece with the rest of Catholicism (the authority of certain figures, the interpretation of other sacraments, and so forth); therefore, Petra's other religious beliefs will point toward her interpretation of the Holy Communion over others. This means that if she sets aside *just this one* belief, she will still conclude that her situation is *impermissive*.

At this point, one might suggest that all Petra needs to do is set aside more of her beliefs. Instead of just setting aside her belief about *R*, she should set aside *all* of her religious beliefs. If she follows this strategy—call it the Expanded Setting-Aside Strategy—she will recognize that there are many religious worldviews compatible with the epistemic rules that she endorses. More broadly, then, we might suggest the following setting-aside test: *My belief B is but one of many permissible alternatives iff it belongs to some collection of beliefs B+ such that, setting aside B+, my epistemic rules do not recommend B over the alternatives.*

The Expanded Setting-Aside Strategy, however, overgeneralizes in an unacceptable way: it delivers the conclusion that *all* of our beliefs are permissive. This is because for any belief or collection of beliefs *B*, there will always be *some* collection of beliefs *B+* such that you recognize the following: setting aside *B+*, the belief-forming methods that you endorse do not privilege *B* over the alternatives. In the limiting case, *B+* is the set of *all* our beliefs. If we set aside all of our beliefs at once, we would recognize that there is nothing to recommend them over other total belief states we could have had—skepticism, counter-inductivism, and obscure conspiracy theories would all be permissible.

But this conclusion is false: there *are* some rational requirements, and *not everything* is permissible. So the Expanded Setting-Aside Strategy is not a good one.

Neither possibility for accommodating acknowledged permissivism looks good. The first possibility (Personal Rules permissivism) only worked by eliminating an important part of the Endorsement Argument. The second possibility, the Setting-Aside Strategy, works only if we are willing to accept an implausible view about acknowledged permissive cases: either that they do not exist at all, that they are very rare, or that *every* case is permissive. Some permissivists might, of course, be happy to take one of these options and accept the Endorsement Argument. But most mainstream permissivists will not want to do this. For these permissivists, there is still work to be done.

II.3. A Question about Consistency Requirements. So far I have worked under the assumption that Schoenfield's view can allow probabilism (or some similar coherence constraint) as a rational requirement. This

seems to be both plausible on its face and the obvious and charitable reading of Schoenfield.[23] Probabilism is also a popular candidate for a rational requirement, which many permissivists will want to defend. However, it is not entirely clear to me how we can accept probabilism, given the Endorsement Argument as it is currently stated. So before moving on, I want to briefly look at how a defense of probabilism might go. Doing so will open up some questions for permissivists who wish to accept the Endorsement Argument: how exactly should we formally characterize this notion, and under what conditions do we endorse certain rules or belief states?[24]

The most obvious route to defending probabilism, of course, is to say that probabilism is one of the rules or methods that we endorse. ("We," here, means just those of us who think probabilism is rationally required.) But *is* probabilism something that we endorse? Recall Schoenfield's definition of endorsement: to endorse a set of cognitive properties is to prefer, when our only goal is accuracy, any cognitive system that instantiates all of those properties to a cognitive system that lacks some of these properties.[25] Schoenfield suggests that if the agent's credences are probabilistic, we might understand endorsement in terms of expected accuracy. Under this interpretation, to prefer a set of cognitive properties C is to be such that, for any cognitive system S, $EA(S|S \text{ satisfies } C) > EA(S|\sim(S \text{ satisfies } C))$.[26]

We can now ask what it would mean to endorse probabilism itself, given this more precise understanding of endorsement. If probabilism is just one of several rules or methods that an agent endorses, it must be the case that the following inequality holds:

$EA(S|S \text{ satisfies conditions } n, n+1\ldots \text{ and probabilism}) >$

$EA(S|S \text{ satisfies conditions } n, n+1\ldots \text{ but not probabilism})$

[23] Throughout the paper, Schoenfield suggests ways to understand her claims in terms of expected accuracy (see much of section 3, as well as note 4, note 16, and elsewhere), which presupposes that an agent's credences are probabilistic. This suggests that she intends her view to at least be compatible with probabilism as a requirement of rationality.

[24] I thank an anonymous referee for helpful comments on this point, which inspired this section of the paper.

[25] *Ibid.*, p. 3.

[26] *Ibid.*, p. 3, note 4. Note that since expected accuracy is assessed relative to a probability function, it seems that an agent can only endorse cognitive properties in this sense if she herself has some credences. This is a realistic assumption for ourselves, of course, but it seems to rule out the possibility that someone could endorse *only* probabilism, or *only* some other consistency constraint. If such an agent had no credences, expected accuracy would not be defined for her. But if she did have some credences, then she would endorse those credences, thereby contradicting the assumption that she only endorsed probabilism. In order to make room for such an agent we would need a different way to understand accuracy-directed preferences.

That is, among the options left open by the *rest* of what one endorses, the probabilistic options have, on average, higher expected accuracy than the nonprobabilistic options.

It is an interesting question why and whether this inequality would be true. It is not at all obvious to me that it would be true given what *I* endorse. Given one way of reading the formal statement above, in a toy case where one's credence in H and in $\sim H$ are both completely unconstrained, and anything between 0 and 1 is permissible, it seems that the inequality fails. This means that given this conditional expected-accuracy interpretation of endorsement, many of us probably do not endorse probabilism.[27]

[27] Suppose we interpret this statement, "$EA(S|S \text{ satisfies } C) > EA(S|\sim(S \text{ satisfies } C))$," as saying the following: for an arbitrary cognitive system S, the expected accuracy of S conditional on S satisfying C is higher than the expected accuracy of S conditional on S not satisfying C. In other words, C-satisfying cognitive systems do better on average than those that do not satisfy C. (I will consider another interpretation in a minute.) Now let us consider a toy case in which one's credence in P, as well as one's credence in $\sim P$, is unconstrained by the rules or methods that one endorses. Let "Pair" be an arbitrary pair of credences consisting of a credence in P and a credence in $\sim P$. We want to know: is $EA(\text{Pair}|\text{Pair sums to } 1) > EA(\text{Pair}|\sim(\text{Pair sums to } 1))$?

Let us use "x" to denote your credence in P, and "y" to denote your credence in $\sim P$, without assuming that these sum to 1. Without loss of generality, suppose P is true. (I am assuming, as is plausible, that accuracy measurements do not depend on which world we are in.) Then, using the Brier score, we can define the inaccuracy of a pair of credences (x, y) as follows: $\text{Inaccuracy}(x, y) = (1-x)^2 + y^2$.

Suppose we pick an arbitrary pair of credences from a uniform distribution, not assuming this pair is probabilistically coherent. We can find the expected accuracy of this pair of credences by taking the integral of our inaccuracy function, between 0 and 1 for both x and y.

$$\int_0^1 \int_0^1 \left[(1-x)^2 + y^2\right] dx dy = \frac{2}{3}$$

The average value of our inaccuracy function for all possible (x, y) pairs is $\frac{2}{3}$.

If (x, y) is probabilistically coherent, $x + y = 1$. So the inaccuracy of this pair of credences will be as follows: $\text{Inaccuracy}(y, 1-y) = (1-(1-y))^2 + y^2 = 2y^2$. Now suppose we pick an arbitrary pair of coherent credences—credences such that x and y sum to 1—from a uniform distribution. To find the expected accuracy of this pair of credences, we can again take the integral:

$$\int_0^1 2y^2 dy = \frac{2}{3}$$

The expected value of a probabilistic pair of credences is exactly as good as the expected value of a nonprobabilistic pair of credences. (Note that this argument built in a couple of assumptions, such as our choosing the pair of credences from a uniform distribution; one avenue for permissivists to respond might involve arguing that this toy case should be set up differently. I will leave these possibilities aside for now, as my main point is to show that defending probabilism via endorsement is not as straightforward as we might wish.)

An anonymous referee drew my attention to another way we could understand endorsement. Rather than spelling it out in terms of conditional expected accuracy, as I do above, we could adopt this stronger interpretation: *if an agent endorses some set of cognitive properties C, then for all S, S′ such that S satisfies C and S′ does not, EA(S) > EA(S′).* This means that all C-satisfying cognitive systems are better, expected-accuracy-wise, than all others.

One option here is to expand our understanding of endorsement, so as to include dominance reasoning or perhaps other ways of valuing or pursuing accuracy. This would allow us to make use of, for example, dominance arguments for probabilism from Joyce and others.[28] This strategy is certainly open to a defender of the Endorsement Argument, but it points to a further line of questioning: exactly what does it take to count as purely having concern for accuracy, and what count as legitimate ways of pursuing accuracy? And does it make sense to rely on expected-accuracy reasoning at one time and dominance reasoning at another—or should we do away with the expected-accuracy understanding altogether?[29] I will not attempt to develop this response in detail but leave it as a question for permissivists who are interested in adopting the Endorsement Argument. These permissivists should not take it for granted that they will end up endorsing probabilism.

A second strategy for vindicating probabilism (or, again, similar coherence requirements) might come not from the content of the rules endorsed but from the *form* of endorsement itself. Certain formal requirements do seem to fall out of the nature of endorsement. For example, suppose rule A recommends credence .5 in P, and rule B recommends credence .6 in P (under the very same circumstances). Is it possible to endorse both rules A and B? It seems that no single system *could* instantiate both rules simultaneously, so it is not possible to endorse a set of rules that contains both A and B. It is also not possible to simultaneously endorse a system that contains A (but not B) and a system that contains B (but not A), since this pattern of endorsement would require incoherent preferences. Perhaps with further development, an argument like this could be extended to show that anything an agent endorses must be probabilistic. To argue this way, we would need to show that our preferences would need to be inconsistent (or something along these lines) if we endorsed, for example, rules that simultaneously recommended .5 credence in P and .6 credence in $\sim P$.

This second strategy also shows some promise. Notice that if we take this strategy, we are changing the role of probabilism in the resulting view. Probabilism would no longer be a *rational requirement* itself but a necessary property of rules that we judge to be rational.

For all I have said here, either of these strategies could turn out to be a successful way of defending probabilism. So I do not take the discussion in this section as an objection to the Endorsement Argument, or to the permissive strategy that it supports; rather, I take it as a call for further clarification of the notion of Endorsement. Which strategy permissivists pick could also have interesting consequences for questions about what sorts of rules or belief states we actually endorse, as well as questions about what sorts of rules or belief states it is even possible to endorse. For now, we can just notice that probabilism does not obviously come out of the Endorsement Argument as stated so far: it seems that defending it will require either reinterpreting or modifying the argument.

II.4. Summing Up. So far we have examined a permissive strategy for connecting rationality and truth. This strategy says that since rationality *only gets us so far*, we should only expect a weak connection between rationality and truth. We have also looked at one particular strategy for establishing that weak connection: Schoenfield's Endorsement Argument. That argument succeeds in linking rationality and truth. However, as I have argued, mainstream permissivists—in particular, those who hold that multiple, precise credences can sometimes be rational in response to permissive bodies of evidence—have reason to reject the Endorsement Argument. First, the Endorsement Argument has the consequence that rational agents should sometimes expect some *irrational* responses to their own total evidence to be more accurate than some *rational* responses. Second, the Endorsement Argument seems to be incompatible with a plausible view on acknowledged permissive cases, according to which such cases are widespread and include religious, political, and scientific disagreement. Finally, I raised a question: how should we understand endorsement, and what would it take for us to endorse coherence requirements like probabilism?

I leave it as an open question what will be the full consequences of adopting this reading rather than the other. However, notice that this interpretation of endorsement rules out situations like Ruth's, which I discussed in the last section. Ruth cannot regard irrational Adam as more (expectedly) accurate than rational Roma, given this stronger interpretation of endorsement. This seems to me to be a reason for permissivists not to adopt the stronger interpretation: we would have to give up some initially plausible views about what permissive requirements could look like. For instance, we could not endorse both the Principal Principle (which will yield impermissive requirements, if we learn the objective chances) and also a permissive view about, say, responding to visual or testimonial evidence. Giving up this possibility is a big cost for mainstream permissivists. (Of course, another way permissivists could respond to Ruth's situation is to say that, contrary to my assumptions in section II.1, *acknowledged* permissive cases are not possible. But this is a big cost too, as I argued in section II.2.) So permissivists should not rush to accept this alternative interpretation of endorsement.

[28] For example, Joyce, "Accuracy and Coherence," *op. cit.*

[29] One might think that it makes sense to rely on dominance reasoning (or Maximin, Minimax, and so on) in situations where we have no credences and hence cannot make expected-accuracy calculations, and to rely on expected accuracy in cases where we have credences. But that argument would not help us here. In this case (see note 27, above) we do get a result from expected-accuracy calculations, but it is just not the right result.

But however limited the Endorsement Argument may be, perhaps it is the best we can do. Can we do better? I will turn to that question in the next section.

III. CAN IMPERMISSIVISM DO BETTER?

In this section I will turn to a way in which *impermissivists* can connect rationality and truth. The impermissivist strategy I will discuss is in some respects simpler than the permissivist strategy. It relies on Claim 1, from before:

> CLAIM 1: When rationality tells us what to believe, usually what it tells us to believe is true.

However, it rejects Claim 2:

> CLAIM 2: Rationality often does not tell us what to believe.

For impermissivism, Claim 2 is false.[30] So if impermissivists can establish Claim 1, they will have a strong connection between rationality and truth.

I will start by looking at an impermissivist argument for Claim 1 from my 2014 paper, "Immoderately Rational."[31] I will then discuss a line of objection to this argument. Fortunately for the impermissivist, I will argue that the objection can be met; impermissivists do have a viable route to Claim 1.

In "Immoderately Rational," I set out an argument for Claim 1, given from the point of view of an agent who is rational according to an impermissive view of rationality. The basic idea is similar to the Endorsement Argument, in that it connects the rules or epistemic methods that a rational agent regards as truth-conducive with those she regards as rational. However, the connection between the two does not rely on any sort of metaepistemological view about the nature of rationality judgments. Instead, it uses the thesis of impermissivism itself. There I argued that because of Immodesty, a rational agent will expect her own epistemic rules or methods to lead to the truth. If she knows that rationality is impermissive, she will know that her own methods are the only rational ones out there. So, she will expect rationality itself to lead to the truth.

Here is the argument as previously presented:

> IMPERMISSIVIST ARGUMENT:
>
> *Where E is any body of total evidence, and C is any credence function:*

[30] Here I am ignoring cases like liar sentences, where the rational requirements (on an impermissivist view) might be indeterminate. I assume that such cases are rare.

[31] Horowitz, "Immoderately Rational," *op. cit.*

> P1. If C is any rationally permissible response to E, then my epistemic rule will recommend C, given E.
>
> P2. If my epistemic rule recommends C, given E, then C maximizes expected accuracy given E.
>
> C. If C is a rationally permissible response to E, then C maximizes expected accuracy given E.[32]

As we can see, the conclusion of this argument is a version of Claim 1: it says that rationality maximizes expected accuracy. Let us go through the premises, seeing what is required for them to be true, as well as what is required for the person giving the argument to know that they are true.

P1 is *true* for the person giving this argument—call her "Irene"—if Irene is rational and rationality is impermissive. Therefore, Irene's own epistemic rule is just the (unique) rational epistemic rule. Irene can *know* P1 if she knows that she is rational and that rationality is impermissive.

P2 is true if we accept an additional assumption: rational agents will be immodest. Again, this means that rational agents will take their own beliefs to have the best prospects for accuracy. How will Irene *know* P2? According to my argument in "Immoderately Rational," to know P2, Irene must know what rationality requires in every circumstance.[33] But it is not very plausible that Irene could come to have knowledge of P2 in this way. (In fact, Schoenfield rejects this impermissivist argument for precisely this reason: she holds the view that we can sometimes be rationally uncertain about what it is rational to believe.[34]) However, there is another possibility for coming to know P2: Irene might simply know that rationality requires Immodesty. Since (as we have already said) Irene knows that she is

[32] This is taken from Horowitz, "Immoderately Rational," *op. cit.*, pp. 46–47. The discussion that follows, however, goes beyond and in some cases disagrees with that original presentation of the argument. In that paper I did not come out strongly in favor of impermissivism over "extreme permissivism," for which I offered a different type of argument. For our present purposes, I will focus on the argument for impermissivism.

[33] *Ibid.*, p. 46. This could happen; maybe the rational requirements are a priori, and since Irene is ideally rational, she knows them all. She could then go through all possible situations one by one, like this:

> "If E is e_1, then rationality requires c_1." [mental calculation] "c_1 maximizes expected accuracy given e_1. If E is e_2, then rationality requires c_2. . ."

This calculation will work out, of course, for the same reason that P2 is true: Irene is ideally rational and immodest. So if rationality requires c_1, then Irene's epistemic rule also requires c_1; and if Irene's epistemic rule requires c_1, then she will take c_1 to maximize expected accuracy. If there were a finite number of possible bodies of evidence, Irene could come to know P2 by surveying all of them in this manner.

[34] Schoenfield, "Permissivism and the Value of Rationality," *op. cit.*, p. 2.

rational, she therefore can also know that she is immodest. This can get her directly to P2.

We have now seen what it takes to get this impermissivist argument off the ground. Impermissivism must be true; Immodesty must be a rational requirement; and the argument must be given from the point of view of a rational agent who knows these things *and* knows that she is rational.

IV. AN OBJECTION TO THE IMPERMISSIVIST ARGUMENT

Let us grant for the moment that rationality is impermissive and that rational agents will be immodest. If both of these are true, it is plausible enough that a rational agent could come to know them a priori (for instance, by doing some epistemology). And let us imagine that the agent giving this argument is rational. How could she come to know that she is rational? This is not so clear. It is certainly not a priori; whether any given person is rational is an empirical fact about the world, and so whether a person knows she is rational depends on what evidence she has. Furthermore, most of us have good evidence to *doubt* that we are rational. So, one might object to the impermissivist argument on these grounds.[35]

In this section I will look at a couple of different ways in which this objection might unfold. Articulating the objection will help us better understand the impermissivist argument. In the next section I will argue that given a new interpretation of the argument, the objection fails.

IV.1. We Cannot Make This Argument for Ourselves. We are now considering how one might object to the impermissivist argument, on the grounds that it requires the person giving the argument to know that she is rational. As a first pass, one might spell out the objection as follows: "If I am going to believe an argument's conclusion on the basis of its premises, I had better believe the premises. I do not believe that I am ideally rational, and so I do not believe P1 of the impermissivist's argument. Therefore I reject the argument."

This first objector has a point. For instance, consider:

BANANA ARGUMENT

P1. I am hungry.

P2. When I am hungry, it is a good idea for me to eat a banana.

C. It is a good idea for me to eat a banana.

[35] This is slightly different from Schoenfield's objection to the impermissivist argument. Schoenfield argued that it is implausible to claim that rational agents can always know what rationality requires. However, as discussed above, the impermissivist argument does not rely on this implausible claim.

This is a valid argument, but I just had lunch. I do not believe P1. So it would be silly for *me* to accept the Banana Argument's conclusion on the basis of its premises. (It does no good to insist: "But the argument is given from the point of view of someone, 'Irene,' who is hungry!" This will not prompt me to accept the conclusion as it applies to *me*.) Is the impermissivist argument like the Banana Argument? Maybe: most of *us* do not know that we are rational, so we do not know the impermissivist's first premise. If the impermissivist argument is one that we are supposed to make for ourselves, coming to the conclusion on the basis of premises that we believe, then it fails.

However, the fact that *some* people do not know or believe certain premises is not the kiss of death for an argument. Consider this argument:

APPLE ARGUMENT

P1. We had five apples this weekend.

P2. We made applesauce, using four apples (and have not obtained or lost any more apples).

P3. $5 - 4 = 1$

C. There is only one apple left.

This is a good argument, and the premises are even true. But my two-year-old son does not accept it. It is not the *argument's* fault that he does not accept it—it is just that my son does not know how to subtract. If he were ideally rational (and informed about this weekend's applesauce project), arguably, he would believe the premises.

IV.2. We Might Not Be Able to Make the Argument, Even If We Were Rational. What if we do not accept the impermissivist argument not because its premises are false, as in the Banana Argument, but because it is like the Apple Argument, and we are (in some respects) like epistemically unsophisticated toddlers? Here is a new hypothesis, then: if we were ideally rational, then not only would the argument's premises be *true* of us, but we would *believe* them, too.

This is certainly a possibility for the impermissivist. She could argue that we are rationally required to believe the premises of the Impermissivist's Argument, and so we are also rationally required to believe the conclusion. However, as previously discussed, this response requires the impermissivist to defend a strong and implausible view to the effect that rationality requires *knowing that one is rational*, which is likely false.

IV.3. Who Cares What This Person Thinks? In light of these first two objections, one might ask: "So who is supposed to deliver this impermissivist argument, anyway? We have established that it is not me, and it is

not necessarily a rational version of me, either. We can imagine a fictional character, Irene, who has the knowledge required to give the argument. But Irene knows things that we do not know and cannot be expected to know. Who cares about Irene, and who cares about her argument?"

We have now landed on what I take to be the most powerful objection to the impermissivist argument: it must be given from a particular perspective, and that perspective requires particular empirical knowledge. But the reply to it, I will suggest, gives the impermissivist a way out. Roughly, the impermissivist should reply that Irene is an (imagined) expert, to whom we should defer. Seeing the argument this way helps us understand why it makes sense to require that Irene know that she is rational.

V. DEFENDING THE IMPERMISSIVIST ARGUMENT

I will argue here that we can accept the impermissivist argument if we interpret it as a case of expert deference. I will argue that Irene—an agent who knows certain a priori truths about rationality *and* knows that she is rational—is someone to whom we should defer. So if we know that an agent like Irene can argue for a certain thesis, we should accept that thesis.

Why should we defer to someone like Irene? It is easy to see why we would want to defer to someone who knows various important a priori truths—such as, in this case, that rationality is impermissive and requires Immodesty. But to give the argument, Irene must also know that she is rational. The rationale for this argument is harder to see.

To provide that argument, let us back up and consider a different question, which might at first seem unrelated. Should we defer to experts when we have more information than they do, regarding the question at hand?[36] For example, imagine that we have an expert meteorologist at our disposal to ask questions about the weather. We ask her whether it will rain an hour from now, and she gives us her answer based on all the latest models and projections. Her answer is: it is very unlikely. Should we believe her? Intuitively, yes. But now add this detail to the story: the meteorologist is working in a windowless room and does not see that storm clouds are approaching from the west. We can see out the window. Now should we believe her? No! What we *should* do is tell her about the storm clouds, and allow her to add that information to her body of evidence—or perhaps we should go back and ask about her *conditional* credence in rain, given that there

[36] As will immediately become clear, my argument here follows Adam Elga, "The Puzzle of the Unmarked Clock and the New Rational Reflection Principle," *Philosophical Studies*, CLXIV, 1 (May 2013): 130–47.

are storm clouds approaching. Only then should we defer to her prediction. The general lesson: if we are going to defer to experts, we should make sure that they do not lack relevant information that we have.

Now let us ask another question: should we defer to experts *who do not know that they are experts*? Adam Elga argues that we should not: the fact that somebody is an expert is, plausibly, a relevant piece of information that we have when we are deferring to that person. We have this information if we are deferring (presumably, this is why we are deferring). And it is relevant because having the information affects what we believe. Just as information about storm clouds outside might change one's rational credences about the weather, information about one's own expertise might change one's rational credences in all sorts of things. So, Elga argues, we should only defer to experts who share our knowledge that they are experts.[37]

We can use Elga's insight to explain how we should interpret the impermissivist argument. The impermissivist argument is not one that we can, or should be able to, make on our own: it is one made by an expert to whom we should defer. The person making the argument, by assumption, knows some relevant a priori truths about rationality (that it is impermissive and that it requires Immodesty). This person also knows that she is rational. We build in this latter piece of knowledge *not* because it is rationally required that the agent know it, but in order to make sure the agent is trustworthy.

This interpretation allows the impermissivist to present her argument without making the controversial assumption that rationality requires us to know that we are rational. What the impermissivist should say, instead, is this: if impermissivism is true *and* we defer to a trustworthy, rational expert's view of the value of rationality, we can come to accept a strong connection between rationality and truth.

The most obvious apparent problem with the impermissivist argument is now taken care of. But one might worry that casting the

[37] Or alternatively: we should only defer to an expert's conditional credences, conditional on the proposition that she is an expert. Here is the argument from Elga (*ibid.*, pp. 10–11):

> Consider [a panel of purported experts] named Cassandra, Merlin, and Sherlock. Conditional on Sherlock being the true expert, what credences should you have? It is tempting to answer: the ones that Sherlock has. . . .But that answer is not correct. For Sherlock himself might be uncertain who is the true expert. And conditional on Sherlock being the true expert, you should not be uncertain who the true expert is. . . .[Y]our credences, conditional on Sherlock being the true expert, should equal Sherlock's credences conditional on Sherlock being the true expert.

Elga uses this argument to motivate a principle of deference to rationality itself, which he calls "New Rational Reflection." He models this on Hall's "New Principal Principle."

argument in terms of expert deference creates a new problem. It is easy to understand why we should accept a valid argument that we make for ourselves, or that we *should* be able to make for ourselves. But what reason do we have to trust this imagined rational agent? What reasoning might lead *us*, with the beliefs that we actually have, to accept someone else's conclusions?

I propose that the impermissivist answer this objection by adding a rational deference or level-bridging principle to her view. A natural candidate would be something along the lines of Elga's New Rational Reflection (though I will not defend it here): our credences should match our expectation of the rational credences, conditional on those credences being rational.

New Rational Reflection: $P(H|P' \text{ is ideal}) = P'(H|P' \text{ is ideal})$[38]

If something like New Rational Reflection is true, then it is true that we should defer to agents like Irene. So if we can show that Irene would believe P on a priori grounds, conditional on the fact of her own expertise, we have a good argument that *we* should believe P as well. This is precisely what the Impermissivist Argument does: it is an argument that Irene can make on a priori grounds, given the assumption that she is rational. So we should believe the conclusion of the argument.

I will conclude this section with a final observation about this impermissivist argument, cast as a case of deference. There might seem to be something peculiar about accepting the conclusion of this particular argument, given that it is put in terms of expected accuracy. Expected accuracy is assessed relative to a particular probability function—in this case, Irene's. Irene's credences are different from ours, since she is ideally rational and we are not. So how can we accept a conclusion that is assessed in terms of Irene's credences?

This question brings up a more general issue about how to defer to experts when what they are telling us has probabilistic content. Such content is always assessed relative to some probability function (or functions), so if this is a genuine problem, we should expect it to arise in many contexts besides this one. I will not attempt to get into this issue here, but I will mention some reasons to think that the problem is not intractable: it is plausible that we can explain deference in this case

[38] *Ibid.*, p. 11. In this context one might worry that such a principle begs the question against permissivism, since if more than one credence function is ideal, New Rational Reflection leads to incoherence. But we can fix this by specifying that the right deference principle applies in impermissive cases only. Since most permissivists agree that *some* cases are impermissive, and since the present argument assumes impermissivism anyway, such a restriction should not cause a problem.

using whatever theory of probabilistic deference turns out to be true. One possibility is to say that in asserting "Probably P" (and by extension, making assertions about expected value or expected accuracy), one is making a recommendation that one's audience adjust their beliefs in such a way as to also endorse "Probably P."[39] Deferring to an expert on this type of view would just amount to taking her recommendation. A related possibility, defended by Sarah Moss, says that when we regard someone as an expert, we take her to know the contents of her assertions (including probabilistic contents). So if we take this expert to know something like "Probably P," we can infer "Probably P" for ourselves.[40] Building on one of these approaches, we could develop a rational deference principle that tells us exactly what deference amounts to in the present context.

For now, I will remain neutral on what exactly it means to defer to Irene's conclusion about expected accuracy in this case. For now the upshot is: the impermissivist argument is best understood as a case of expert deference. If impermissivism is true, then an expert—someone to whom we should defer—can conclude that rationality maximizes expected accuracy. This gets the impermissivist a strong connection between rationality and truth.

VI. CONCLUSION

We began by asking whether rationality can be a guide to the truth and *also* allow some leeway in what we can believe. We then looked at two arguments purporting to draw a connection between rationality and truth: one available to permissivists and one available to impermissivists. As we have seen, the impermissivist argument establishes a stronger connection between rationality and truth than the permissivist argument. However, the *strength* of the truth-connection is not all that matters. In closing I will discuss another difference between the two arguments and summarize what I take to be the state of the debate.

This difference appears to favor the permissivist argument, at least initially. That is: the permissivist argument is pitched at *you*, the reader, and (if it works) takes *you* to its conclusion, mostly relying on premises you already believe. The impermissivist argument, on the other hand, requires you to buy into a complete "package deal" in order to reach

[39] See Eric Swanson, "The Application of Constraint Semantics to the Language of Subjective Uncertainty," *Journal of Philosophical Logic*, XLV, 2 (April 2016): 121–46, for an expressivist view that supports this thought; and see Matthew Mandelkern, "How to Do Things with Modals," forthcoming in *Mind and Language*, for a contextualist view with similar consequences.

[40] Sarah Moss, *Probabilistic Knowledge* (Oxford: Oxford University Press, 2018). See especially section 5.4.

its conclusion: you must accept impermissivism, *and* Immodesty, *and* a rational deference principle. Because of this, it looks like the permissivist argument will be more dialectically effective than the impermissivist argument. Because dialectical effectiveness is so clearly a goal of Schoenfield's Endorsement Argument, we should spend some time discussing it directly. Although the Endorsement Argument is designed for dialectical effectiveness, in the end I do not think this consideration favors it over the impermissivist argument.

One reason is that dialectical effectiveness is not a very important virtue of an argument. In philosophy we are not *just* trying to convince one another. And a very broad demand for dialectical effectiveness—say, one that says we need to convince not only the rational or mostly rational, but also toddlers and other people with serious rational pathologies—would be both impossible to meet and useless to aim for. Even an audience only of mostly rational adults is hard to target: what will convince one overall-reasonable person may not convince another at all.

Still, let us focus on *you*, and assume that you are more or less reasonable. Whether an argument is dialectically effective for *you* will depend on whether you accept its premises. A little less obviously, dialectical effectiveness depends on whether you accept the *conclusions* that the argument commits you to (one reader's modus ponens is another's modus tollens, and so forth). I do not know you or what you believe. But I predict that if you are a mainstream impermissivist, you will be happy with the impermissivist argument discussed here. Its commitments (impermissivism, Immodesty, and a rational deference principle) are ones you should be happy to accept. If you are a mainstream permissivist, however, I predict that you will not be happy with the Endorsement Argument. This argument most crucially relied on a specific view about the nature of rationality judgments, which does not sit well with "Personal Rules," or intrapersonally impermissive, versions of permissivism. And the Endorsement Argument was also incompatible with plausible views about acknowledged permissive cases. In this respect, the endorsement argument does not seem to do so well on dialectical effectiveness after all.

To sum up: mainstream impermissive views can explain how rationality is connected to truth. But the truth-connection provided by the Endorsement Argument comes at a steep price for mainstream permissivists. For permissivism, the truth problem remains unsolved.

SOPHIE HOROWITZ

University of Massachusetts, Amherst

CAUSAL DECISION THEORY AND DECISION INSTABILITY*

In the face of new information, your deliberation about what to do may change course. Sometimes this can happen when your only new information comes from your deliberation itself and from the course it has taken so far. Sometimes deliberation's path follows turns and switchbacks, and a stable decision may be hard to find. One of the best-known illustrations of this, the story of the man who met death in Damascus, appeared in the infancy of the subjective theory of rational choice known as *causal decision theory*. Causal decision theory and treatments of decision instability have long been linked.[1] Causal decision theory is much discussed presently, and objections to it come in several forms, old and new. The theory is in excellent health, but that is not my present topic. Here I will explore the use of causal decision theory, when you deliberate about what to do in *Death in Damascus* and in similar decision problems.

A straightforward and general understanding of the scope of causal decision theory is presented here. We can call it *unadorned* causal decision theory, but really, *causal decision theory* is already a fine term. When it is applied to problems like Death in Damascus, we will find that the interplay of your rational assessments and your rational beliefs during deliberation is a fascinating topic, but not a source of difficulty for causal decision theory.

In what follows, I use the Death in Damascus problem to illustrate decision instability and *deliberation dynamics*. Then I consider a purported counterexample to causal decision theory, representative of others, namely Andy Egan's *Murder Lesion*

*Thanks to Shyam Nair, Brian Skyrms, Simon Huttegger, Peter Vanderschraaf, Steven Reynolds, and especially to Jim Joyce for helpful discussions of this paper. My views about topics discussed here have benefited from many conversations and exchanges with others. Particular thanks to William Harper, Christopher Hitchcock, Paul Weirich, Alan Hajek, Susan Vineberg, Danny Hintze, Jaemin Jung, and Wes Anderson. Special thanks to Jim Joyce, and most of all, to Brian Skyrms.

[1] Causal decision theory and the story of the man who met death in Damascus were both introduced in Allan Gibbard and William Harper, "Counterfactuals and Two Kinds of Expected Utility," in C. A. Hooker, J. J. Leach, and E. F. McClennen, eds., *Foundations and Applications of Decision Theory, Volume I: Theoretical Foundations* (Dordrecht, the Netherlands: D. Reidel, 1978), pp. 125–62; reprinted in W. L. Harper, G. A. Pearce, and R. Stalnaker, eds., *Ifs* (Dordrecht, the Netherlands: D. Reidel, 1981), pp. 153–90. Brian Skyrms's example of the *mean demon* shares the form of Death in Damascus, and it appears in his early discussion of deliberation dynamics in "Causal Decision Theory," this JOURNAL, LXXIX, 11 (November 1982): 695–711.

0022-362X/19/1605/263–277

problem.[2] A simple response on behalf of causal decision theory, called the *Simple Response*, shows how Murder Lesion and similar problems fail to be counterexamples, and it clarifies the general use of the theory in problems of decision instability. I then compare unadorned causal decision theory and the Simple Response to previous treatments by Frank Arntzenius and by Jim Joyce.[3] There are differences among the three, and I recommend the unadorned theory and the Simple Response. But there is also much agreement among them, particularly in the practical import of adopting them. The present effort does not press on to consider other objections to causal decision theory, but it makes room for better discussions of their merits.

I. DECISION INSTABILITY AND DELIBERATION DYNAMICS

Allan Gibbard and William Harper considered the issue of stability in rational choice, illustrated by the example of the man who met death in Damascus:

> Consider the story of the man who met death in Damascus. Death looked surprised, but then recovered his ghastly composure and said, 'I am coming for you tomorrow'. The terrified man that night bought a camel and rode to Aleppo. The next day, death knocked on the door of the room where he was hiding and said, 'I have come for you'.
>
> 'But I thought you would be looking for me in Damascus', said the man.
>
> 'Not at all', said death 'that is why I was surprised to see you yesterday. I knew that today I was to find you in Aleppo'.
>
> Now suppose the man knows the following. Death works from an appointment book which states the time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo.[4]

[2] Andy Egan, "Some Counterexamples to Causal Decision Theory," *Philosophical Review*, CXVI, 1 (January 2007): 93–114.

[3] Frank Arntzenius, "No Regrets, or: Edith Piaf Revamps Decision Theory," *Erkenntnis*, LXVIII, 2 (March 2008): 277–97. James M. Joyce, "Regret and Instability in Causal Decision Theory," *Synthese*, CLXXXVII, 1 (July 2012): 123–45.

[4] Gibbard and Harper, "Counterfactuals and Two Kinds of Expected Utility," *op. cit.*, pp. 185–86. The story is a variant of one told by W. Somerset Maugham in *Sheppey* (1933) and alluded to in the title of John O'Hara's *Appointment in Samarra* (1934). Other versions of the story are far older, dating to the ninth century and probably earlier; see "When Death Came to Baghdad," in Idries Shah, ed., *Tales of the Dervishes* (London: Jonathan Cape, 1967). Different cities appear in various earlier versions of the story.

To evaluate his options with causal decision theory, the man uses states of the world that form a *K*-partition. Each such state is one he takes to be outside his causal influence and sufficiently specific given his interests. Let K_D and K_A be 'Damascus is inscribed' and 'Aleppo is inscribed'; they form a *K*-partition. If, for example, he believes it is Damascus rather than Aleppo, and his belief $pr(K_D)$ is greater than ½, then causal decision theory endorses going to Aleppo.[5] But that decision seems unstable: when the man comes to believe he is about to go to Aleppo, he has new information that influences his other beliefs, including his beliefs about what is inscribed, about K_D and K_A. Since he takes death's appointment book to be based on reliable predictions, his anticipation that he will go to Aleppo raises his belief that Aleppo is inscribed after all, so $pr_n(K_A)$ is greater than ½, and $pr_n(K_D)$ is less, which makes Damascus the better option. But when the man anticipates he is about to choose Damascus, that newer information again influences his beliefs, so $pr_{nn}(K_D)$ exceeds ½, which makes Aleppo the better option..., and so on.

What should the man do?[6] The sense of instability in problems like *Death in Damascus* arises when he can reevaluate his options in light of information arising from his deliberations. This is a good setting for the theory of deliberation dynamics, where updates of your beliefs inform your continuing deliberations, leading to new assessments of your options that in turn provide reasons for further belief updating.

[5] Here I use the *K*-expectation version of causal decision theory due to Skyrms, "Causal Decision Theory," *op. cit.* Causal decision theory endorses going to Aleppo in the sense that going to Aleppo's expected utility U is maximal, and U represents his preferences. To say that causal decision theory endorses action A is to rely on a principle that endorses actions that maximize U. I follow other discussions in accepting that principle; my concern here is the proper application of the principle in extended deliberation. To say that causal decision theory endorses acting at a particular moment t, or when one is in a particular epistemic state e, is to rely on another principle. More about that below.

[6] One plausible answer is toss a coin, or adopt some internal method of randomizing his choice, thereby pursuing a mixed strategy; see William Harper, "Mixed Strategies and Ratifiability in Causal Decision Theory," *Erkenntnis*, XXIV, 1 (January 1986): 25–36. Neither pure act (remain in Damascus, go to Aleppo) is ratifiable, but a 50-50 mixture of those acts is. A good idea, perhaps, but suppose that mixed strategies are ruled out as viable options—if the man were to use one, death would know, and would interrupt his appointment-keeping to find the man wherever he is, as in Paul Weirich, "Decision Instability," *Australasian Journal of Philosophy*, LXIII, 4 (December 1985): 465–72. Fanciful examples aside, problems that forbid or penalize mixed acts thereby impose a restrictive exogenous constraint on the decision-maker's options. Having said that, however, we should be careful about "solving" a decision problem by altering it with additional options and offering a solution to the revised problem.

The theory can be applied to deliberations about many sorts of decision problems, simple and complex.[7]

Deliberation dynamics applies to deliberation that takes place over time.[8] At each moment, your beliefs about states of the world (for example, Damascus is inscribed) underwrite your current assessments of your options. Those assessments conform, let us suppose, to subjective rational decision theory; throughout this discussion, causal decision theory is the theory in use. If we assume that the values you attach to outcomes (for example, life, death) are not shifting, then when your beliefs are stable, so are your assessments of your options. But as in the case of the man who met death, your beliefs may not be stable. At a given moment, your current assessments may give you reason to change your beliefs about what you will do, and to change your beliefs about states of the world that matter to the outcome of doing it. The new beliefs underwrite new assessments, and we can entertain the trajectories of your repeatedly revised beliefs and assessments over time. Sometimes those trajectories may display oscillations, as we imagined in the case of the man who met death.[9]

The trajectory of your beliefs will depend on the details of your dynamics. How do your beliefs about what you will do depend on the values you attribute to each option?[10] You recognize at t that one action A looks better than its alternative B, that $U_t(A) > U_t(B)$; perhaps you also recognize the difference between them, or the ratio $U_t(A) / U_t(B)$, on some particular U_t scale. How much does that lead you to increase your belief at t_+ that you will do A, your $pr_{t+}(A)$? The answers to such questions throughout your deliberation might be given by a dynamical rule. Many such rules are possible; among them are rules that *seek the good*, according to which you raise your probabilities that you will perform actions exactly when you currently regard those actions as better than their alternatives, or more precisely, as better than the *status quo*, which is your current expectation of the outcome of the problem you are deliberating about.[11] If at some moment t your beliefs lead you to regard your options as equally good, so that $U_t(A) = U_t(B)$, then your assessments give you no reason, under dynamics that seek the good, to alter those beliefs and assessments. You are at an *equilibrium* of the dynamics. Since you then regard each action as equally good, to choose one or the other is to break a tie, or "pick" among the tied options.

Returning to Death in Damascus, then, suppose that during his deliberation, the man is attentive to his evaluations of his options, and that they inform his beliefs about what he will soon do as well as what is inscribed in Death's appointment book. Suppose that at time t_1 during his deliberation, he regards going to Aleppo as the better action, so that $U_{t_1}(A) > U_{t_1}(D)$, and he realizes that he does; he then raises his belief that he will go to Aleppo, $pr_{t_2}(A) > pr_{t_2}(D)$, and also that death will be waiting for him there, $pr_{t_2}(K_A) > pr_{t_2}(K_D)$.[12] Then, when he reevaluates his options at t_2 with those new beliefs, he sees Damascus as the better action, $U_{t_2}(D) > U_{t_2}(A)$, which gives him reason to revise his beliefs again. Under plausible assumptions, in the Death in Damascus problem deliberation that seeks the good will eventually lead the man's beliefs to a stable equilibrium, where he sees neither act as better than the other.[13] At that point, his tied evaluations give him no

[7] See Skyrms, "Causal Decision Theory," *op. cit.*, pp. 701–06. In later work, Skyrms developed an important connection between dynamic deliberation and well-founded solution concepts for noncooperative games among Bayesian players. See Brian Skyrms, *The Dynamics of Rational Deliberation* (Cambridge, MA: Harvard University Press, 1990). In decision problems with more than two options, deliberation can be significantly more complex than in Death in Damascus. Arntzenius, "No Regrets, or: Edith Piaf Revamps Decision Theory," *op. cit.*, and Joyce, "Regret and Instability in Causal Decision Theory," *op. cit.*, each invoke Skyrms's deliberation dynamics in their treatments of Death in Damascus and similar problems, as we will see.

[8] This includes intermittent deliberations: the offer expires on the day after tomorrow; the election is on Tuesday; a plan must be in place next week; I have one hour to make a move; and so on. When deliberation takes time, further news may arrive and sway its course. But effects of outside news are ignored here.

[9] When you deliberate about different sorts of problems, your changing beliefs and evaluations follow different sorts of paths. When causal decision theory is used to evaluate your options in a *Newcomb Problem*, for example, deliberation may yield straightforward convergence to high confidence that you will take both boxes, and that the opaque box will be empty, since an increasing confidence that you will take both boxes does not lead you to think it would be better to do otherwise.

[10] Also, how *much* do your beliefs about what you will do depend on your assessments of those values? We might explore the idea that your beliefs respond to other influences too, but here I leave that for another occasion, and suppose that the belief changes are driven only by your shifting assessments of your options.

[11] The idea of dynamics that seek the good is Skyrms's; see *Dynamics*, *op. cit.*, p. 30. Such dynamics must also raise the sum of the probabilities of all the actions better than the *status quo*. Both Arntzenius and Joyce require that dynamics for rational agents seek the good; see Arntzenius, "No Regrets, or: Edith Piaf Revamps Decision Theory," *op. cit.*, p. 293, and Joyce, "Regret and Instability in Causal Decision Theory," *op. cit.*, pp. 132–33.

[12] The dynamic rule expresses the change in his beliefs about what he will do, $pr_{t_2}(A)$ and $pr_{t_2}(D)$. Accompanying changes in his beliefs about what is inscribed, $pr_{t_2}(K_A)$ and $pr_{t_2}(K_D)$, satisfy Jeffrey-conditionalization if his conditional probabilities such as $pr(K_A/A)$ are stable and there are no further complexities.

[13] In general, given sufficient time, we can expect convergence to equilibrium from reasonable starting points. For Death in Damascus, continuous deliberation dynamics that seek the good are guaranteed to converge to equilibrium, but they will not display the oscillations I have described. Discrete-time dynamics that seek the good may well display oscillations; with plausible properties such as a dampening in the learning over time, convergence to equilibrium can be guaranteed. See Skyrms, *Dynamics*, *op. cit.*; and William Harper, "Decision Dynamics and Rational Choice," forthcoming in Billy Dunaway and David Plunkett, eds., *Meaning, Decision, and Norms: Themes from the Work of Allan Gibbard* (Ann Arbor, MI: Maize Books). See also Greg Lauro and Simon Huttegger, "Decision Dependence and Causal Decision Theory," manuscript.

reason to further adjust the beliefs that underlie them. At the equilibrium state in the original Death in Damascus problem, $pr_{eq}(K_A)$ and $pr_{eq}(K_D)$ are both ½, as are $pr_{eq}(A)$ and $pr_{eq}(D)$. His action will be the outcome of some way of dealing with the tie between *A* and *D*. A general feature of equilibrium states, whether your deliberation leads you to them or not, is that you see your available options as equally choiceworthy, as having equal expected utility. It may also happen that you then believe that you are as likely to do one act as the other, but that need not be so in problems that lack the symmetry of Death in Damascus.

Why should the man embark on this deliberative journey? There is at least this reason: a rational choice should be based on all of your relevant beliefs at the time you make it. So, *if* you believe at time *t* that death is more likely to go to Aleppo than to Damascus, $pr_t(K_A) > pr_t(K_D)$, your evaluations at *t* of your options, $U_t(A)$ and $U_t(D)$, must use those beliefs. Or, to put it another way, a rational decision theory, such as causal decision theory, is properly used only when those evaluations do so. Is it incumbent upon you to possess such beliefs in the midst of deliberation? We will return to that question soon.

The original version of Death in Damascus is a symmetric problem, but asymmetric versions are easily given; just add an incentive against travel that makes the outcomes of staying in Damascus a little better than the corresponding outcomes of traveling to Aleppo.[14] Or, imagine that death's appointment book more reliably predicts the traveler's presence when he is in one city than when he is in the other.

Decision instability has become prominent in work on causal decision theory. One reason is that it displays the wider scope of the theory, beyond problems where causal-dominance reasoning applies. Another is that problems displaying instability have been offered as *counterexamples* to causal decision theory.

II. MURDER LESION AND THE SIMPLE RESPONSE

Andy Egan challenged causal decision theory with a set of examples that he judged to be counterexamples to the theory, and his challenge has received wide attention. One of the examples is the *Murder Lesion* problem:

> Mary is debating whether to shoot her rival, Alfred. If she shoots and hits [*S* & *H*], things will be very good for her. If she shoots and misses [*S* & *M*], things will be very bad. (Alfred always finds out about unsuccessful assassination attempts, and he is sensitive about such things.) If she doesn't shoot [~*S*], things will go on in the usual, okay-but-not-great kind of way. Though Mary is fairly confident that she will not actually shoot. . .she thinks that it is very likely that if she were to shoot, then she would hit [$S \mathbin{\square\!\!\rightarrow} H$]. So far, so good. But Mary also knows that there is a certain sort of brain lesion that tends to cause both murder attempts and bad aim at the critical moment. If she has this lesion, all of her training will do her no good—her hand is almost certain to shake as she squeezes the trigger. Happily for most of us, but not so happily for Mary, most shooters have this lesion, and so most shooters miss. Should Mary shoot?[15]

Following Egan, let the utility of shooting and hitting (*S* & *H*) be 10, the utility of shooting and missing (*S* & *M*) be –10, and the utility of not shooting (~*S*) be 0 throughout.[16] Mary's initial beliefs are that she is unlikely to shoot, $pr_i(S) < .5$. She also thinks that if she did, she would hit, $pr_i(S \mathbin{\square\!\!\rightarrow} H) > .5$. Her belief in that causal conditional is dependent on whether or not she shoots, since shooting is correlated with having the lesion; so $pr_i(S \mathbin{\square\!\!\rightarrow} H/\ S) < .5$. However, her initial unconditional belief in that conditional is high, as just specified, since she initially thinks *S* is unlikely.

With those initial beliefs, a causal decision theory calculation will yield $U_i(S) > U_i(\sim S) = 0$, since the better outcome of *S*, namely *S* & *H*, is weighted by the high probability $pr_i(S \mathbin{\square\!\!\rightarrow} H)$, while the worse outcome *S* & *M* is weighted by the low probability $pr_i(S \mathbin{\square\!\!\rightarrow} M)$. So causal decision theory endorses shooting. Egan regards that as a flawed endorsement:

> It's irrational for Mary to shoot. . . .In general, when you are faced with a choice of two options, it's irrational to choose the one that you confidently expect will cause the worse outcome. Causal decision theory endorses shooting. . ..In general, causal decision theory endorses, in these kinds of cases, an irrational policy of performing the action that one confidently expects will cause the worse outcome. The correct theory of rational decision will not endorse irrational actions or policies. So causal decision theory is not the correct theory of rational decision.[17]

[14] Reed Richter, "Rationality Revisited," *Australasian Journal of Philosophy*, LXII, 4 (December 1984): 392–403.

[15] Andy Egan, "Some Counterexamples to Causal Decision Theory," *Philosophical Review*, CXVI, 1 (January 2007): 93–114, at p. 97. Notation added.

[16] There is symmetry in these payoffs, but perhaps not in the beliefs: *most* shooters have the lesion, but whether the same proportion of non-shooters lack it is unsaid; it is *nearly certain* that those with the lesion miss; in light of her training, Mary thinks it is *very likely* that she would hit. Nothing I say here depends upon the problem being as symmetric as Death in Damascus. In what follows, we might identify states of a *K*-partition (have the lesion, do not have the lesion) and use the *K*-expectation version of causal decision theory, as before. But here I follow Egan, who uses causal conditionals as in the Gibbard–Harper version of causal decision theory.

[17] Egan, "Some Counterexamples to Causal Decision Theory," *op. cit.*, pp. 97–98. Phrases referring to a different example have been omitted.

The act of shooting is intuitively irrational, Egan says, and widely judged to be so.

> . . .we have (or at least my informants and I have) clear intuitions that it's irrational to shoot or to press, and rational to refrain in *The Murder Lesion*. . .[18]

There is a response to Egan's view of the example. Egan's case for the irrationality of causal decision theory's endorsement (that Mary shoot) is the intuitive irrationality of shooting. No basis for the intuition is offered, but it is not hard to feel, nor hard to explain. What is happening? Mary begins with the beliefs that she lacks the lesion and that shooting would be effective; based on those beliefs causal decision theory endorsed shooting.[19] That is the right endorsement given her beliefs and values at that time. But in preferring shooting, she probably has the lesion and will very likely miss. So Mary comes to confidently expect, and we who contemplate her problem come to confidently expect, that shooting will cause the worse outcome. That is what the first step in the deliberative process tells her. What is Egan's intuition, if not the result of taking that step? At that point, however, when Mary has *that* belief, causal decision theory endorses *refraining*; that is what Mary's current utilities will tell her. Egan applies the endorsement that causal decision theory makes at one time (shoot) to a decision at a later time, after Mary's beliefs have changed, and he sees a flaw where there is none. The error is in the supposition that causal decision theory is forever committed to its endorsement under Mary's initial beliefs.

> The resulting theory enjoins us to *do whatever has the best expected outcome, holding fixed our initial views about the likely causal structure of the world.* The following examples show that these two principles come apart, and that where they do, causal decision theory endorses irrational courses of action.[20]

What causal decision theory really endorses is what has the best expected outcome, given our *current* views about the likely causal

[18] *Ibid.*, p. 98.

[19] It is worth remarking that users of causal decision theory are no more prone to murder, premature death, disease, or psycho-killing than anyone else. The window dressings of our examples should be more varied; outcomes might be prizes or penalties, large or small, rather than death. Many *games* exhibit instability: Battle-of-the-Sexes-with-a-Twin, for example. The theoretical issues apply to small stakes as well as large, a point worth remembering when deliberation has a cost.

[20] Egan, "Some Counterexamples to Causal Decision Theory," *op. cit.*, p. 96. Emphasis Egan's.

structure of the world.[21] This applies to us in the first person, as deliberators (use our current beliefs), and in the third person, as judges of what causal decision theory says about others (use their current beliefs). Egan's argument suffers from a mistake about what causal decision theory endorses.[22]

A second issue is that an intuition that refraining is uniquely rational has doubtful reliability. We are invited to deliberate a little bit, but not very far, about what to do, and to stop the deliberation at an arbitrary point, with no motivation given for stopping there. If Mary correctly assesses her options at that point, when she thinks she has the lesion, refraining is better. But that provides her reason to think that, as a refrainer, she likely lacks the lesion, and that belief makes shooting the better option after all (according to her). . .and so on. Even if the mistake about causal decision theory's endorsements were absent, the example would at best indicate a problem with joining causal decision theory to a special unmotivated assumption about when deliberation must end. It establishes no problem for causal decision theory, which is consistently a good guide to rational action. So says this line of response; let us call it the *Simple Response.*

What, then, does causal decision theory endorse in the *Murder Lesion* problem? If you have Mary's initial beliefs and deliberate no further, it endorses shooting. If you have a different initial belief, that you probably have the lesion, it endorses refraining. If you have access to and appreciation of your deliberative states, and your beliefs and

[21] Egan actually states the principle correctly later in his paper in a different context (*ibid.*, p. 102), but it is clear that he relies on the incorrect version throughout the paper. Without it, what is the purported counterexample?

[22] I say that the result of the first deliberative step in Murder Lesion explains Egan's intuition that refraining is rational and shooting is irrational. But accounting for intuitions about particular examples is an uncertain project, and there may also be other intuitions in play. One might be uneasy about instability itself, for example. Why use a theory that makes an endorsement, then retracts it, then reinstates it, and so on? The retraction suggests that the endorsement should not have been made in the first place. But the second recommendation is not a retraction of the first endorsement, it is an update, in light of new information. It is one thing to have an advisor who wavers in his advice for no discernable reason, another thing when the advice changes as he receives a stream of relevant information. At some point, though, you will stop listening. In any case, it is hard to see how an aversion to instability points to the rationality of one specific option rather than the other, how it points to refraining as rational, and shooting as not. Egan's intuition, and his informants', appears to be different. Joyce suggests that framing and loss aversion may be at work, and that could be so; see Joyce, "Regret and Instability," *op. cit.*, p. 135. But Egan's example emphasizes what Mary comes to confidently expect, rather than her great aversion to the worse outcome. . .I think the best explanation of the intuition Egan describes is the one given with the Simple Response. Whether or not that is so, the error about what causal decision theory endorses, when your rational beliefs shift during your deliberation, remains. Thanks to an anonymous referee for raising questions concerning intuitions about these problems.

assessments interact in extended deliberation, causal decision theory makes a succession of endorsements at each stage of your self-reflecting dynamical deliberation. The endorsements may change, but each one is correct for your beliefs and values at the moment it is made. Your deliberation may end in a variety of ways. You may get tired, you may have other things to do, the world may interrupt, or you may reach equilibrium. If your deliberation ends in action, the rational action to perform is the one endorsed by your current beliefs, when deliberation ended. Recall that if you reach equilibrium, you see your options as equally worthy, and you shoot or refrain by dealing with the tie. There is no single action, for every deliberator, that use of causal decision theory requires or leads to. So goes the Simple Response to Murder Lesion and to similar purported counterexamples. And in general, when deliberation is unstable and wavers between options, so says unadorned causal decision theory.

III. DOES THE EQUILIBRIUM RULE?

According to the preceding account, causal decision theory delivers assessments of your options at each moment of your deliberative process, whether or not you are at an equilibrium. In the company of a principle that endorses actions with maximal causal expected utility U, at each moment of your deliberative process, unadorned causal decision theory endorses the action or actions that currently have maximal U. Further constraints on use of causal decision theory are neither imposed nor needed. In particular, no requirement is made that you must reach an equilibrium, or that you must look to an equilibrium prior to reaching it, in order for causal decision theory to endorse one, or some, of your options. Of course, if you do reach stable equilibrium, causal decision theory will deliver stable evaluations of your options, and will equally endorse those tied with maximal U_{eq}.[23]

Contrasting accounts given separately by Frank Arntzenius and Jim Joyce disagree. Each of their accounts offers an adornment to causal

[23] Since application of causal decision theory does not depend on your arriving at equilibrium, the issue of whether your starting point and your dynamics will lead you to equilibrium is interesting, but not crucial to using the theory. This idea, and more, is also expressed by Skyrms, in *Dynamics*, *op. cit.*, pp. 36–37: "It is possible—perhaps likely—that deliberation will have reached an equilibrium by the moment of truth, in which case her decision will be a best response. On the other hand, in the absence of special knowledge, it is no more likely if the moment of truth arrives before equilibrium that she will make a worse response rather than a better one. The present expected utilities just calculated may not be the ones which will obtain at the moment of truth, but they are in a sense the decisionmaker's best estimate of them." In earlier work, Weirich considers, but too quickly rejects, the idea that causal decision theory applies at each moment of extended deliberation. See Weirich, "Decision Instability," *op. cit.*, pp. 466–67.

decision theory, and each argues that the equilibrium state is the unique perspective from which to ascertain your rational action. In a nutshell, Arntzenius's view is that, in decision problems like those we are considering, you should be in the equilibrium state when making your decision. Joyce's view is that epistemic rationality will lead your deliberation to the equilibrium state, and it is only then that you are in a satisfactory epistemic position to make assessments that should guide your choice. Joyce would say that what causal decision theory endorses all along, even before you reach equilibrium, is what it endorses then, when your epistemic state is satisfactory and you rank each viable option equally.

Let us look briefly at Arntzenius's treatment first. It is developed in the company of his suggestion for understanding how you might perform a mixed strategy, an option that is a probabilistic mixture of your pure options (go to Aleppo, remain in Damascus). Arntzenius associates mixed acts with states of belief; to act when your rational belief that you will go to Aleppo, $pr(A)$, is x and your belief that you will remain in Damascus, $pr(D)$, is $1-x$, is to perform the mixed act $(xA, (1-x)D)$.[24] After showing that deliberation that seeks the good will eventually arrive at the equilibrium state, Arntzenius asks,

> Must one really model a rational person as a deliberator who changes his credences during the deliberation? No, one need not. Indeed it is a little bit awkward to do so. After all, if one is ideally rational, then how could there be any stage at which one has the 'wrong' credences?[25]

His second question might suggest sympathy for a treatment like the one I am advocating, but Arntzenius instead recommends that we set aside non-equilibrium beliefs:

> So, as long as we are idealizing, let us simply say that a rational person must always be in a state of deliberational equilibrium. The dynamical model of deliberation that I gave can be taken to merely amount to a crutch to make us realize that there always exists a state of deliberational equilibrium.[26]

[24] Performance of the mixed act yields one or the other of the pure options. The connection between your equilibrium beliefs and the option you take raises interesting questions; I save them for another occasion.

[25] Arntzenius, "No Regrets," *op. cit.*, p. 294.

[26] *Ibid.* Arntzenius argues for the existence of the equilibrium under continuous dynamics that seek the good. The decision problem he explicitly considers, *Psycho Johnny*, is based on one of Andy Egan's examples. Johnny has two options, and its complexity is similar to that of Death in Damascus. More generally, in problems with more than two available options, the existence of stable equilibria under a given dynamical rule is a complex issue and not always guaranteed. See Lauro and Huttegger, "Decision Dependence," *op. cit.*, section 4.