oxford studies in epistemology | volume 2

OXFORD STUDIES IN EPISTEMOLOGY

OXFORD STUDIES IN EPISTEMOLOGY

OXFORD STUDIES IN EPISTEMOLOGY

Volume 2

Edited by

Tamar Szabó Gendler and John Hawthorne

EDITORS' PREFACE

It is with great pleasure that we introduce the second issue of *Oxford Studies in Epistemology*, inaugurated in 2005 as a member of Oxford's expanded collection of *Oxford Studies* serials. The diverse set of essays that appear here represent some of the most interesting epistemological work going on in the English-speaking world today, providing the reader with a glimpse into an active and vibrant area of philosophical investigation.

Published biennially under the guidance of a distinguished editorial board, each issue of *Oxford Studies in Epistemology* seeks to include an assortment of exemplary papers in epistemology, broadly construed. *OSE* aims to publish not only traditional works in epistemology, but also work that brings new perspectives to traditional epistemological questions, and that opens new avenues of investigation.

These commitments are evident in the contents of the second issue. The papers that appear here are diverse in their foci, but uniform in their quality. Two are concerned with the question of how the challenge of radical skepticism can be met, with Richard Fumerton raising challenges for one line of thought in ''Epistemic Conservatism: Theft or Honest Toil'' and Nico Silins providing a defense of another in ''Basic Justification and the Moorean Response to the Skeptic.'' Two others explore tensions that surround the idea of epistemic rationality: David Christensen's ''Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals'' and Allan Gibbard's ''Rational Credence and the Value of Truth.'' Gibbard's paper serves as a target for responses by Frank Arntzenius in ''Rationality and Self-Confidence'' and Eric Swanson in ''A Note on Gibbard's Rational Credence and the Value of Truth,'' with a reply by Gibbard in ''Aiming at Truth over Time: Reply to Arntzenius and Swanson.'' Finally, two essays reflect the journal's ongoing commitment to bringing work on epistemology in related fields to the attention of philosophers: linguist Kai von Fintel and philosopher Anthony Gillies's ''An Opinionated Guide to Epistemic

Modality'' and primatologist Laurie Santos's ''The Evolution of Irrationality: Insights from Non-human Primates.''

As in the past, some of the papers that appear here were brought to our attention by members of the editorial board, others were solicited directly from authors; all were refereed by the members of our Editorial Advisory Board, to whom we are grateful. Thanks are due to all of its members: Stewart Cohen (Arizona State University), Keith DeRose (Yale University), Richard Fumerton (University of Iowa), Alvin Goldman (Rutgers University), Alan Hájek (Australian National University), Gil Harman (Princeton University), Frank Jackson (Australian National University and Princeton University), Jim Joyce (University of Michigan), Scott Sturgeon (Birkbeck College London), Jonathan Vogel (University of California at Davis), and Tim Williamson (University of Oxford) We are also indebted to our outstanding managing editor Roald Nashi, for his superb editorial assistance, and to Peter Momtchiloff, for his continuing support of this project.

<div align="right">

Tamar Szabó Gendler, Yale University
John Hawthorne, Oxford University

</div>

CONTENTS

LIST OF FIGURES

LIST OF CONTRIBUTORS

Frank Arntzenius
UNIVERISTY OF OXFORD

David Christensen
BROWN UNIVERSITY

Kai von Fintel
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Richard Fumerton
UNIVERSITY OF IOWA

Allan Gibbard
UNIVERSITY OF MICHIGAN

Anthony Gillies
UNIVERSITY OF MICHIGAN

Laurie Santos
YALE UNIVERSITY

Nicholas Silins
CORNELL UNIVERSITY

Eric Swanson
UNIVERSITY OF MICHIGAN

This page intentionally left blank

PAPERS

This page intentionally left blank

1. Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals

David Christensen

Formally inclined epistemologists often theorize about ideally rational agents—agents who exemplify rational ideals, such as probabilistic coherence, that human beings could never fully realize. This approach can be defended against the well-known worry that abstracting from human cognitive imperfections deprives the approach of interest. But a different worry arises when we ask what an ideal agent should *believe* about her own cognitive perfection (even an agent who is in fact cognitively perfect might, it would seem, be uncertain of this fact). Consideration of this question reveals an interesting feature of the structure of our epistemic ideals: for agents with limited information, our epistemic ideals turn out to conflict with one another. This suggests that we must revise the way we see ideal agents in epistemic theorizing.

1. IDEAL VERSUS HUMAN-CENTRIC RATIONALITY

What would an ideally rational agent believe? Of course, the answer depends on just what kind of ideally rational agent is in question. But when epistemologists consider this question, they don't simply answer ''everything true''. Rationality, after all, involves reacting correctly to the evidence one has, but does not seem to require having all possible evidence about everything. Thus, if we seek to

understand rationality by constructing a model of ideally rational belief, we will not concentrate on an omniscient being. Instead, we'll consider a non-omniscient thinker who nevertheless is in certain respects cognitively perfect. We might, for example, stipulate the following kinds of things about such an ideally rational agent's beliefs. They would not be based in wishful thinking. They would be independent of the agent's likes and dislikes. They would respect whatever evidence the agent had. And they would respect the logical relations among claims the agent had beliefs about. Let us call a non-omniscient agent who nevertheless is ideally rational an IRA.[1]

This general approach to theorizing about rationality dovetails nicely with the tradition which relates rationality to thinking logically, and then characterizes rational belief with the aid of formal logic. Those who see belief as a binary, all-or-nothing, kind of state have thus often taken logical consistency and logical closure to be rational ideals. And those who conceive of beliefs as coming in degrees have taken conditions based on probabilistic coherence—which can be seen as little more than applying standard deductive logic to graded beliefs—as ideals.[2]

Of course, this whole formal approach to thinking about rationality has been criticized. The main line of criticism takes off from the fact that ideals such as logical consistency or probabilistic coherence are very clearly far beyond the capacities of any human to achieve—even more so than complete freedom from prejudice or wishful thinking. Such ideals require, for instance, that an agent believe (or, in the case of coherence conditions, be completely certain of) every logical truth. Why then, it is asked, should rules that might apply to a peculiar sort of imaginary beings—ideal thinkers with limited information—have any bearing on us? The fact that we

---

[1] For a sense of how widespread this approach is, see the following *Stanford Encyclopedia* entries: James Joyce on Bayes' Theorem, Sven Ove Hansson on Logic of Belief Revision, James Hawthorne on Inductive Logic, Robert Koons on Defeasible Reasoning, and William Talbott on Bayesian Epistemology. Books in this tradition include Savage (1954), Ellis (1979), Horwich (1982), Maher (1993), and Levi (1997).

[2] By ''conditions based on probabilistic coherence,'' I mean not only conditions requiring agents to have precise real-valued degrees of confidence satisfying the laws of probability, but also less restrictive conditions modeling rational degrees of belief by sets of probability functions, or qualitative probabilities. For convenience, I'll use the term ''probabilistic coherence'' to refer to this whole family of conditions.

humans have the particular limitations we do, it is urged, is not just some trivial footnote to epistemology; it's a central aspect of our epistemic predicament. Interesting epistemology—epistemology for humans—must take account of this fact.[3]

I think that this line of criticism should be resisted. While there are certainly some projects in epistemology that must take careful account of human limitations, they do not exhaust interesting epistemology.[4] For example, if one's epistemological project were to characterize our ordinary, casual way of using the words ''rational'' and ''irrational'' to apply to people, then it might be hard to see how humanly unattainable ideals would play an important role: everyone fails to live up to humanly unattainable ideals, but we obviously don't call everyone ''irrational''. But there's little reason to think that epistemology should be restricted to such a thin notion of rationality. (Similarly, ethics should not be restricted to studying moral ideals that are perfectly attained by the ordinary people we'd hesitate to call ''immoral''.)

A related point applies to the project of developing a notion of rationality that's closely linked to an ''ought''-implies-''can'' notion of epistemic responsibility. Clearly, we don't want to blame anyone for failing to live up to an unattainable ideal. But there are certainly evaluative notions that are not subject to ''ought''-implies-''can''. I would argue that our ordinary notion of rationality is one of them: when we call a paranoid schizophrenic ''irrational'', we in no sense imply that he has the ability to do better.[5]

Another epistemic enterprise in which the importance of highly idealized models might be questioned is the so-called ''meliorative project''—epistemology aimed at our cognitive improvement. Some have claimed that any interesting epistemology must be aimed at providing us with guidance to help ourselves (or perhaps others) to think better. I personally doubt that philosophers are particularly well-equipped for this sort of endeavor. But even putting that doubt aside, I see no reason to think that the sole point

---

[3] For some representative instances of this line of criticism, see Hacking (1967), Cherniak (1986), Goldman (1986), Kitcher (1992), and Foley (1993).

[4] I cannot make the case for this claim here in full. What follows is a brief sketch, with references to more sustained discussions.

[5] See Feldman and Conee (1985), Alston (1985), and Christensen (2004: 6.4) for more discussion of this point.

of epistemology should be the production of manuals for cognitive self-help.[6]

What projects are there, then, which make manifestly unattainable epistemic ideals worth studying? One such project is that of assessing us as a species. After all, there is no reason to suppose—even if we are the cognitive cream of the mammalian crop—that we're the be-all and end-all of any evaluative epistemic notion we come up with. Indexing epistemic perfection to the cognitive capacities of *homo sapiens* clearly begs some interesting questions.

But the most important reason for resisting the impatience some express about idealized models of rationality does not depend on the interest of evaluating humans as a species. It is clear that our ordinary rationality judgments are based in assessments of people's levels of performance along certain dimensions of epistemic functioning. And these dimensions may well be ones whose extremes are beyond human reach. Freedom from wishful thinking is a plausible example. Predicting consequences of social policies in a way that's untainted by self-interest is another. More examples include evaluating other people's behavior and character without prejudice from emotional ties, or from bigotry based on race or sexual orientation. And a natural candidate for this list is having beliefs that do not violate logic.

If rationality consists (at least partly) in good performance along this sort of dimension, then one natural approach to understanding rationality more clearly is to study candidates for rationality-making qualities by abstracting away from human cognitive limitations, and considering idealized agents who can perfectly exemplify the qualities under consideration. Is logical consistency of all-or-nothing belief a rational desideratum? What about probabilistic coherence of degrees of confidence? How should agents update their beliefs when presented with new evidence? It seems that questions like these may be approached, at least in part, by asking ourselves, ''What would an IRA believe?''

Now it is important to see that the suggestion here is not that questions about rational ideals *reduce to* questions about what ideal agents would believe. Any such reduction would likely run afoul of immediate counterexamples involving, for example, beliefs about

---

[6] See Christensen (2004: 6.5) for further references and discussion.

the existence of ideally rational agents.[7] It might be the case that any ideally rational agent would be quite confident that there were conscious beings who could not remember making any cognitive errors; this does nothing to show that such a belief is rationally mandatory in general. But this sort of problem does not, I think, undermine the usefulness of IRAs in studying rationality. It's just that one has to be alert to the distinction between those aspects of an IRA's beliefs which help make it ideally rational, and those that are mere side-effects of the idealization.

It might be insisted that we must still connect considerations about IRAs with claims about us non-ideal agents. However, there are simple, plausible ways of doing this. For example, one attractive thought is that if the constraints that apply to IRAs describe the endpoint of a spectrum, then the closer an actual agent's beliefs are to that end of the spectrum, the better (presumably, ceteris paribus). Efficiency in cars is a nice analogue here: perfect efficiency is impossible, but (ceteris paribus) the more closely one approaches this end, the better. Moral principles also might work this way: I am undoubtedly psychologically incapable of being perfectly fair or generous; but the more closely I approximate perfect fairness and generosity, ceteris paribus, the better.[8]

To my mind, some of the most promising applications of highly idealized theorizing about rationality involve taking probabilistic coherence as a constraint on degrees of belief. Considerations along the lines rehearsed above, I think, show that some of the most common objections to idealizations involving probabilistic coherence, on the grounds that they abstract so far from human limitations, are misguided. I would like, then, to say something like: ''Well, of course none of us can be probabilistically coherent, but that's no big deal. We can see that coherence is an ideal in part by showing that IRAs have coherent credences. And as far as my own beliefs are concerned, the closer I can come to having coherent credences, the more rational my beliefs will be.''[9]

---

[7] Williamson (2000: 209–10) makes essentially this point.

[8] Zynda (1996) argues along these lines; see also Christensen (2004: 6.5).

[9] This thought presupposes that we can make sense of one's beliefs coming closer to coherence. Zynda (1996) develops a way of making sense of this notion in order to give normative force to the unrealizable ideal of coherence.

Unfortunately, I now think that the claim that IRAs are coherent is probably false, and that the claim about the rationality of approaching coherence in my own beliefs is at least problematic. The reasons for this are related to, but ultimately quite different from, the worries about idealization described above. They raise what seems to me an interestingly different difficulty for the standard way of using ideal agents in theorizing about rationality, a difficulty flowing from the structure of our epistemic ideals.

2. IDEAL RATIONALITY MEETS POSSIBLE COGNITIVE IMPERFECTION

The problem I would like to examine involves a very different way in which cognitive imperfection poses an obstacle to taking probabilistic coherence as a rational ideal. The problem arises from an agent's apparently rational reflection on her own beliefs. Let us begin by thinking about a case involving a clearly non-ideal agent:

Suppose I prove a somewhat complex theorem of logic. I've checked the proof several times, and I'm extremely confident about it. Still, it might seem quite reasonable for me to be somewhat less than 100 percent confident. I should not, for example, bet my house against a nickel that the proof is correct. After all, balancing my checkbook has shown me quite clearly that my going over a demonstrative argument, even repeatedly, is not sure proof against error. Given my thorough checking, my being in error this time may be highly unlikely; nevertheless, it is hard to deny that I should give it some nonzero credence. Let us call the theorem I've proved T. And let us use M to denote the claim that, in believing T, I've come to believe a false claim due to a cognitive mistake. The question now arises: given this sort of doubt, how strongly—ideally speaking—should I believe T?

It seems that my giving some slight credence to M is required by my recognition that I may sometimes exhibit cognitive imperfection. And to the extent that I have any rational credence at all in M, I must have some rational credence in the negation of T (since M obviously entails $\sim$T). So my confidence in T should fall short of absolute certainty; in probabilistic terms, it should be less than 1.

But if something like this is correct, it seems to raise an obstacle to taking coherence as a rational ideal for me—an obstacle quite different from that raised by the fact that coherence is humanly unattainable. For according to this argument, it would not be *rational* for me to have full confidence in T, a truth of logic. In fact, if I did manage to have the coherence-mandated attitude toward T, the argument would urge me to back away from it. So the problem is not the usual one cited in connection with human cognitive limitations. It's not that I *can't* achieve the probabilistically correct attitude toward T—in this case, I may well be perfectly capable of that. The problem is that, in the present case, it seems that my beliefs would be worse—less rational—if I were to adopt the attitude toward T that's mandated by probabilistic coherence.

It is worth pointing out that the problem is not just about having maximal belief in logical truths. To see this, suppose I give some positive credence to $\sim$T. Now consider what credence I should give to ($\sim$T v C), for some ordinary contingent claim C. If it is different from my credence in C, then my credence in these two contingent claims will violate the principle that logically equivalent claims get equal credence. On the other hand, if my credence in ($\sim$T v C) is equal to my credence in C, then I will violate the principle that my credence in a disjunction of logically incompatible disjuncts should be the sum of my credences in the disjuncts.

The basic problem is that coherence puts constraints on my credences based on the logical relations among all the claims in which I have credences—including contingent claims. To the extent that I have doubts about whether certain logical relations hold, and to the extent that those doubts are reflected in my credences, coherence may be violated—even when explicit consideration of logical truths is not involved. For another example, suppose that contingent claim P logically entails contingent claim Q, but I am not absolutely certain of this. In at least some such cases, it would seem that I should then have somewhat higher credence in P than in (P & Q). But if I do, then again I have given logically equivalent contingent claims different levels of credence.

Clearly, this problem should be disconcerting to those of us who would advocate coherence—either the simple version, or one of the standard generalizations—as a component of ideal rationality. To my mind, the threat it poses is significantly deeper than that posed

by the fact that probabilistic perfection is not humanly possible. Thus it's worthwhile seeing whether the one might resist the claim that it would be irrational for me to be coherent.

3. CAN I RATIONALLY BE CERTAIN OF T?

Suppose one were to argue as follows:

> Certainty Argument: Granted, I must give ~T at least as much credence as I give to M. But I have the strongest possible kind of justification for full confidence in T—I've proved it demonstratively. So I should give it full confidence, and should give ~T, and thus M, zero credence. (After all, my proof of T serves as a proof of not-M!) I may not be a perfect being, but I have the best possible reason for believing T, and thus the best possible reasons for being certain that I haven't come to believe a false claim due to a cognitive mistake.

I think that this argument should not tempt us. To see why, suppose that I work out my proof of T after having coffee with my friend Jocko. Palms sweaty with the excitement of logical progress, I check my work several times, and decide that the proof is good. But then a trusted colleague walks in and tells me that Jocko has been surreptitiously slipping a reason-distorting drug into people's coffee—a drug whose effects include a strong propensity to reasoning errors in 99 percent of those who have been dosed (1 percent of the population happen to be immune). He tells me that those who have been impaired do not notice any difficulties with their own cognition—they just make mistakes; indeed, the only change most of them notice is unusually sweaty palms. Here, my reason for doubting my proof, and the truth of T, is much stronger. It seems clear that in the presence of these strong reasons for doubt, it would be highly irrational for me to maintain absolute confidence in T. Yet the certainty argument would, if sound, seem to apply equally to such extreme cases.

Could this verdict possibly be resisted? Could one argue that, initial appearances to the contrary, we actually can embrace the certainty argument, even in the strong doubt case? One way of attempting this would capitalize on distinguishing carefully between two sorts of cases: the bad ones, where the drug has impaired my reasoning and my proof is defective, and the good ones, in which I'm one of the lucky 1 percent who is immune to the drug's effects and my proof is correct. It might be pointed out that we cannot assume that what would be irrational for the person in the bad case would be irrational for someone in the good case. After all, those in the good case have constructed flawless sound proofs of T, and those in the bad case have made errors in reasoning. To say that what holds for one must hold for the other would be to conflate having a correct proof with seeming to oneself to have a correct proof. So it might be argued, that although it would be clearly wrong for most people who find out that they've been dosed to dismiss the resulting doubts, at least if I am in the good case, I am in a different epistemic position, and I may rationally dismiss the doubts.[10]

Now I think that there is something to this point. I would not claim that the epistemic situations of the drug-sensitive person and the immune person are fully symmetrical. After all, the drug-sensitive person in the envisioned type of situation makes a mistake in reasoning even before she finds out about the drug, and the drug-immune person does not. But granting the existence of an asymmetry here does not mean that it is rational for the drug-immune person to disregard the evidence suggesting that he has made an error. And it seems clear—especially when one keeps in mind that those who are affected by the drug don't notice any impairment in their reasoning—that, given the evidence suggesting I've made a mistake, it would be irrational for me to maintain full confidence in my reasoning, even if I happen to be in the good case.[11]

Thus we cannot exploit the real epistemic asymmetry between the drug-sensitive and drug-immune people to argue that the latter may after all avail themselves of the certainty argument. And if this is

---

[10] The envisioned argument is inspired by a point Thomas Kelly (2005) makes in a different context, though he should not be saddled with it here.

[11] See Feldman (manuscript) and Christensen (2007) for discussion of parallel points relating to the epistemology of disagreement.

correct, it is hard to see how we can support applying the Certainty Argument even to the original cases involving mild self-doubts raised by memories of misadventures in checkbook balancing. Nothing in the Certainty Argument hinged on the mildness of the doubt about my proof. In fact, it does not seem that even the weak positive reasons for doubt provided by the checkbook-balancing memories are needed to prove the point. Suppose I've never made a mistake in balancing my checkbook or in any other demonstrative reasoning. Surely that doesn't license me in being certain that such mistakes are impossible. And as long as such mistakes are possible, it is hard to see how I can be *certain* that they have not occurred. Even if my reason for doubt is slight, and, so to speak, metaphysical—so slight that in ordinary cases, I wouldn't bother to think about it—still, it would seem irrational to be absolutely certain that I had not come to believe a false claim due to a cognitive mistake. And thus it would seem irrational for me to be absolutely certain of T.

If this is right, it underlies a troubling result for those of us who see coherence as a rational ideal. For the only way I can live up to the ideal of coherence here would seem to be by irrationally dismissing the possibility that a cognitive mistake led me to believe T falsely. Being certain of logical truths seems not only to be something that I can't always do—it seems like something I often shouldn't do. And that makes it hard to see what kind of an epistemic ideal probabilistic coherence could be.

4. WOULD AN IDEALLY RATIONAL AGENT BE CERTAIN OF HER OWN IDEALITY?

The troubling result flows from the fact that I must believe myself to be epistemically fallible. But if rational ideals can be thought of as those that would make an *ideal agent's* beliefs rational, perhaps this is not the right way to think about the issue. Perhaps an IRA would not only never make a cognitive error, but would also (rationally) be certain of her own cognitive perfection. If that were so, then we could at least hold that an IRA would have probabilistically coherent beliefs. And this might help explain a sense in which coherence was, after all, an epistemic ideal. The idea would be something