Oxford Studies in Epistemology – final version

Aiming at Truth Over Time

Reply to Arntzenius and Swanson

Allan Gibbard
University of Michigan
Ann Arbor, Michigan 48109
U. S. A.

I want to thank both Frank Arntzenius and Eric Swanson for their fine, illuminating commentaries. Both propose that my analysis of belief should be made dynamic. In my paper I considered only a simplest possible case, the static case with a single uncertain proposition and its negation. I might have gone on to consider a more complex thinker, prone to change degrees of credence as new evidence comes in. I agree with Arntzenius and Swanson that the dynamic case needs investigating. I think, however, that for the dynamic case, most of the lessons I drew reappear in new forms.

1. The Problem

In my paper, recall, I tried to make sense of the idea that "belief aims at truth." I considered epistemic rationality, and asked whether it can somehow be explained as answering to a pure concern with truth. By epistemic rationality, I mean rationality in one's degrees of credence. (For short, following David Lewis, I call degrees of belief "credences"). The epistemic rationality of a state of belief is different from its overall desirability. It is not the same thing as rationality in acting to affect the belief state, or the belief state's being the kind one might go for given the choice. The upshot of my inquiry was both negative and positive. Concern for truth, I first argued, might take any of various forms. Some of these are friendly to epistemic rationality, and some are not. In arguing this, I took for granted standard ideas of what epistemic rationality consists in—or at least, I took the standard decision-theoretic conditions as necessary for perfect epistemic rationality. Concern for truth as such, I thought I showed, couldn't explain epistemic rationality. Epistemic rationality answers to a concern for truth only if the concern takes a special form: that of concern with truth for the sake of guidance.[1]

[1] The theorems I appealed to and the core of my argument were, as I indicated, drawn from the work of statistician Mark Schervish.

I helped myself to standard requirements on credence in order to see if they are self-endorsing. Since I took these requirements as assumptions, no argument of the kind I gave could possibly convince anyone of these requirements who wasn't already convinced. Jim Joyce undertakes a more ambitious kind of argument, one that addresses a person who doubts that epistemic rationality requires standard coherence in one's credences—where "standard coherence", as I'm using the term, amounts to satisfying the usual axioms of probability. Joyce tries to show, on the basis of things that such a person would accept, that probabilistic incoherence is defective. His vindication of standard coherence was meant to be non-pragmatic, and one of my conclusions was that a non-pragmatic vindication, along the lines he attempts, is not to be had. I criticized some of the conditions he himself laid down; they aren't all required, I argued, for a person to qualify as purely concerned with truth. Then I helped myself to standard decision theory, parts of which he meant to vindicate, squeezed all I could from the notion of purely epistemic goals, and still couldn't get the main result that he derived from his conditions. It would seem that if a non-pragmatic vindication along Joyce's lines isn't to be had even with assumptions that help themselves to the view to be vindicated, it isn't to be had at all. Epistemic rationality, I concluded, isn't to be explained as what a sheer concern with truth must endorse.

Concern with truth in one special form, though, did seem to do some explaining. That was the lesson I drew from theorems of Mark Schervish. The form is concern with truth on pragmatic grounds, but of one particular kind—or alternatively, a concern with truth that mimics such a pragmatically grounded concern. Attempted pragmatic vindications of probabilism are of course well known, and aspire to be much more general than the limited pragmatic vindication that I ventured. Whether any of these pragmatic vindications work has been widely debated, and Frank Arntzenius may be unconvinced by some of them. Nothing I showed adds anything to those debates. It does seem to be a lesson of my argument, though, that any successful vindication will have to be pragmatic. More guardedly, I should say, that's the lesson unless something more can be squeezed out of notions of truth-conduciveness than I myself could identify.

Both commentators put my puzzle in ways somewhat different from what I intended. According to Azs, my puzzle is that I "can see no good reason to be self-confident," no good reason not to judge my beliefs epistemically deficient. Not exactly: as Arntzenius indicates later on, whatever reasons one has for one's degrees of credence are reasons for thinking them right, and thus that one has got them right. It's just that I don't see how the reason can take a particular form: thinking—even circularly—that one's credences aim at truth in a way that is optimal given one's evidence. I thus don't see how an intrinsic concern for the truth of one's beliefs could in any way underlie epistemic rationality. (Arntzenius,

as I read him, doesn't see how either, though he may think we could see the folly of such approach without any argument like mine.)

I also don't think that it is "rationally acceptable to judge one's own degrees of belief as epistemically deficient." A perfectly rational person, I would think, will not so judge. It may well be rationally acceptable, I said, to wish that one's degrees of belief were different from what they are—even when it's rational to care only about their closeness, by some standard, to full truth. Epistemic deficiency, though, is different from being unwanted. It's different even from aiming badly at truth. My puzzle brings into question not epistemic rationality but a specious way of explaining it. Can we "give a purely epistemic justification for why our belief states are as they are" (if they are ideally rational)? Arntzenius says that I think we can't, but in truth I don't know and I would hope that we could. My conclusion was that we can't give such a justification along the lines that I scrutinized, explaining epistemic rationality as somehow well aimed at the truth for its own sake.

According to Swanson, I think that epistemic rationality constrains credal value functions to be credence-eliciting. Again, not exactly: A person could be epistemically rational and value truth in all sorts of ways. He might even disvalue truth, but find himself epistemically rational against his wishes. The thesis I scrutinize in my paper, once I think I have made sense of it, allows for this. If a person is epistemically rational, goes the thesis, it is *as if* she valued truth for its own sake and could choose her credences at will. Most of us can't choose our degrees of credence at will, and a person who can't might conceivably be epistemic rational to perfection, but wish that she weren't.

One more set of preliminary remarks: Swanson suggests strengthening my characterization of concern with truth. The concern with truth, he says, should be symmetric: one should value credence in the negation of a claim, should it be false, just as one values credence in the claim should it be true. He notes that this makes no difference to the conclusions I drew, but even so, I'll register my disagreement. To be sure, the truth of S amounts to the falsehood of $\neg S$, and so trivially, the truth of S and the falsehood of $\neg S$ are of equal import. It doesn't follow, though, that the truth and falsehood of S are of equal import. Take almost any example: let S be Newtonian physics, or a value for the speed of light, or the new Hair-Brane theory in particle physics. Must uncertainty that S is true in case it is true and uncertainty that S is false in case it is false be equal failings, from the standpoint of a pure, scientific thirst for truth? I don't see why. We're comparing, say, a person who is 95% certain of the inverse square law for gravity when it isn't quite the correct law, with a person who is 95% certain that it isn't the correct law when it is. Why must their high but misplaced confidence and their correct residual doubts be of equal purely epistemic import, when its being precisely true would tell us a lot and its being not quite right would leave it wide open just what is right? I don't

know which residual doubt is more important, but once we're convinced that not every 1% difference in the credence one might have matters equally, why think that these two do matter equally? "Credence aims at accuracy," Swn proposes, and "a valuation of credal accuracy should not arbitrarily privilege credence in truths or credence in falsehoods by valuing them asymmetrically." I agree that a blanket policy of treating all truths one way and all falsehoods another isn't even possible, since the negation of a falsehood is a truth. I suggest, however, that a particular truth and its negation might very well be treated asymmetrically by a person who still rightly counted as valuing truth purely for its own sake.

2. The Dynamic Case: Updating over Time

Both Arntzenius and Swanson analyze thinkers who take in new evidence over time and somehow modify their credences in its light. I agree that such an analysis is needed, and I'll turn first to Swanson's treatment. Swanson argues that a dynamic analysis changes the lesson to be drawn—at least for beings like us, with our limitations. With this I mostly disagree.

Note first a crucial feature of the static case. A coherent believer who wants only truth, recall, may wish that her credences were different from what they are. That was the central point with which I began. Note, though: In that case, if she got what she wanted, she still wouldn't be satisfied. Her credences would be different from what they are, and so her prospective valuations of the various possible arrays of credences one might have would, in this counterfactual case, be different from what they are in actuality.

Swanson's treatment of the dynamic case plays on this feature. He considers hypothetical believers "who always act to maximize expected value" and "can choose their own credences" (10[c]). He shows that "a believer who has non-credence eliciting value functions *and can choose her own credences* engages in some very odd doxastic behavior over time."[2] His dynamic believer, able to choose her credences anew at each updating, ends up with credences she wouldn't have wanted in the first place for the case of receiving the string of evidence that she receives.

This is quite right, as he shows conclusively. The remaining question is how it bears on the claim that epistemically rational credences do in some sense "aim at truth". For the static case, I argued, rational credences do aim at truth, but in a special way: it is as if they aimed at truth for the sake of guidance. Valuing truth in a way that mimics valuing it for the sake of guidance, though, I said, is far from the only way one could value truth for its own sake. Now the way I set up the question for the static case has a parallel for

[2] P. 10 [10f], emphasis mine, and with the pronoun changed to facilitate reference.

the dynamic case. It is this parallel, I'll argue, and not the case that Swanson analyzes, that bears on whether epistemically rational policies for credences and their revision can be explained as aiming at truth.

For the static case, recall, I put the question as whether, if a person is epistemically rational in her credences, it is *as if* she valued truth and had been able to choose her credences at will. (If she valued truth in a way that made her want different credences from the ones she has, we now note, she would want not only to have those different credences, but to lose her power to set her credences at will. Otherwise she would end up, after a series of new choices of credences, with credences different from the ones she now wants.) The answer to my question depends, of course, on what qualifies as "valuing truth"—but I'll put off further discussion of that until later, and assume for now that I was right about what valuing truth in one's credences consists in. Our question now is how to pose the parallel question for the dynamic case. For the dynamic case, we suppose that the believer values truth not only for her credences at the outset, but for the credences she will come to have as new evidence crops up. What she needs to evaluate, then, is whole ways she might be disposed to form credences and update them. In actuality, we are supposing, she is epistemically rational, and so her actual epistemic dispositions, whether she wants them to be that way or not, consist in starting out with a coherent, epistemically rational array of initial credences and then updating by standard conditionalization. The question is whether she will be glad that those are her epistemic dispositions. If she in some way values truth and truth alone, will her actual epistemic dispositions be the ones she most prefers to have?

The answer to this question for the dynamic case exactly parallels the answer for the static case. What are the alternatives among which she can have preferences? As both Swanson and Arntzenius recognize, she isn't restricted to wishing to update by standard conditionalization. Swanson proposes another restriction, though, which I'll accept as an important restriction to explore. Let's confine our consideration to beings who, like us, can't keep track of their past histories of updating. Suppose, indeed, that our believer can't even aspire to more, that she is constrained to wish only for epistemic dispositions that don't require keeping track of such matters as how many times she has updated. On each updating, we require for the world as she wishes it were, she must apply a rule that takes her current credences and the new item of evidence, and on the basis of these alone delivers a revised array of credences. What dispositions, under this restriction, will she most prefer to have? That is our question.

Swanson provides the machinery that delievers an answer to this question. Take the "report relation" R that Swanson defines, which takes actual to wished-for credences. Look, as he shows that we must, for an array of dispositions that mimic updating from

her actual credences by standard conditionalization and then "applying R at the end". Because of the informational restriction, we must now, I agree, further require that her way of valuing truth yields a report relation that is injective (that is, that it is a one-to-one function from the interval $[0, 1]$ onto itself). As he notes, however, this isn't a severe restriction; it allows for many report relations that aren't the identity relation—that aren't the R of a believer who most prefers the epistemic dispositions that she in fact has.

Here are the dispositions she most prefers to have (though so long as R isn't identity, she doesn't in fact have them): the dispositions are, in effect, at each stage as new evidence arrives, to revert to her epistemically rational credences, apply standard conditionalization, and then go to the new credences that, in actuality, she prefers for the case of having that evidence. This works as follows: Let ρ_0 be her actual, epistemically rational credences at time 0, and for discrete times $t = 1, 2, \ldots$, let ρ_t be the credences that, with her actual dispositions, she would have at time t having received a string of evidence $E_1, E_2, \ldots, E_t$. What arrays of credence $\sigma_0, \sigma_1, \ldots, \sigma_t$, we now ask, does she wish she were disposed to have on receiving that string of evidence. She wishes, as Swanson says, that each σ_t were the one she would get by starting out with her actual initial credences ρ_0, updating by standard conditionalization, and applying R at the end. But a non-standard updating rule that she can wish for would accomplish just that. (Indeed it is a rule that Swanson considers, though it doesn't work for the situation that Swanson considers, where the believer is stuck having to wish for states where she could wish further and get what she then wished.) Let her wished-for initial credences σ_0 be the ones that result from applying R to her actual initial credences ρ_0. Let her wished-for dispositions to update be this: that on receiving each new piece of evidence, she update as if she first had reverted to the credences ρ_{t-1} that she is actually disposed to have, then had updated these by standard conditionalization, and finally had applied the report relation R to the result.

This gives her a wished-for updating rule that fits Swanson's restriction on wished-for information. The rule, more fully put, consists in first (i) applying to her wished-for credences σ_{t-1} the inverse R^{-1} of the report relation, yielding her rational credences ρ_{t-1}, then next (ii) applying standard conditionalization C_t, defined as $C_t(\rho_{t-1}) = \rho_{t-1}(\cdot/E_t) = \rho_t$, and finally, (iii) applying the report relation to the result to get $\sigma_t = R(\rho_t)$. Her wished-for updating function is thus RC_tR^{-1}, the transformation that results from applying successively the transformations R^{-1}, C_t, and R. This may be messy, but applying this updating rule would, with enough sheer calculating power, require only keeping track of one's current credences and what the new evidence is.

This dynamic parallel to the static case differs sharply from the case that Swanson analyzes. I examine only what the rational believer who values truth actually wants. Swanson examines a case where the believer, on the arrival of each new piece of evidence,

gets what she wants and so forms new preferences which are then accorded at the next updating. This, as he shows, isn't something to want—unless one wants precisely the initial credences one has. His treatment plays, as I have said, on a feature that the dynamic and the static cases share: that in case the believer isn't satisfied with her credences, if she got what she wants she still wouldn't be satisfied.

Would this feature itself, though, indicate that she doesn't genuinely want truth in her credences? Does it show that she fails really to value truth and truth alone? If it does, then perhaps the dictum that belief aims at truth can still be interpreted as correct. We can still maintain that any rational believer who values truth *genuinely* will be glad she has the credences she does.

But this feature indicates no such thing. All sorts of things we might genuinely value in beliefs will display this feature. The suicide prefers self-inflicted death to his prospects otherwise—but once he kills himself, he no longer has this preference. His preference is none the less genuine. Or take an instance that is more complex: I want comfort, but I also want to be emotionally braced for rude surprises. I want not to be completely terrified all the time, but still to be somewhat prepared for the things I dread. What credences would, on balance, prospectively best meet these and my other competing desiderata? They may not be the credences I actually have and that I regard as epistemically rational. Perhaps, for the sake of comfort, they'd discount the liklihood of some of the things I fear—but still not too much, or I'll be too unprepared if terrifying things do happen. What credences I most want to have will thus depend, among other things, on how likely I now take various nasty eventualities to be. For that reason, if I had the credences I actually most want, the calculations I now make would no longer apply. I'd want even lower credence in fearsome things that might befall me. None of this means, though, that I don't now genuinely value comfort as a benefit that my credences might yield.

I conclude, then, that the dynamic case works like the static one—with a qualification. A being fully coherent in belief and preference might intrinsically value truth and truth alone and still want a credal policy different from her actual ones. In the dynamic case, she might want both different initial credences and a different updating rule. As Swanson indicates, the updating rule she wants will in some cases demand extraordinary amounts of information. Not so, however, in cases where her epistemic preferences yield, in Swanson's terms, a report relation that is injective. Then, the rule she most wants can run on the same information as standard conditionalization: one's credences prior to the new evidence and what the new evidence is. Valuing truth, then, even in this restricted way, needn't lead an ideally rational person to want the credences she has. Epistemically rational credences, then, can't be explained just as being what you'd want if you valued truth and truth alone.

Arntzenius, for the dynamic case, starts out with just the right question. "What should I now regard as epistemically the best policy for updating my degrees of belief in light of the evidence I will get." He shows, for the particular case he considers, that the policy will depart from standard conditionalization as its updating rule. I agree, as I have indicated in my treatment of Swanson. He finds problems with this, however. First, it goes against diachronic Dutch book arguments, and if we lose Dutch book arguments, we have no answer to why credences ought to satisfy the axioms of probability—why, as I'm using the term, they ought to be coherent. Dutch book arguments, though, are pragmatic, not purely epistemic, and I haven't questioned pragmatic arguments for classical decision theory. My point is that we can't get a certain kind of purely epistemic argument to work. As for why to have degrees of belief that satisfy the axioms of probability, that is an excellent question, but not one that I took up. I considered only degrees of belief in a single proposition.

Arntzenius's second problem with dropping standard conditionalization is that one loses "the ability to set one's degrees of belief so as to maximize the current expected epistemic utility of those future degrees of belief." Here what I said about Swanson applies. In the linear case, the one that Arntzenius chiefly analyzes, Swanson's report relation R isn't injective. We can still ask Arntzenius's question of what, by my actual lights, would be my prospectively best updating policy. The policy that looks prospectively best by my initial lights will still look prospectively best over time as new evidence comes in. But the policy will make heavy informational demands; it can't prescribe credences as a function just of what one's credences are before a piece of evidence comes in and what that evidence is. Arntzenius may be suggesting this when he says that if I had my desired credences, "I would lose the information as to what I should do were I to learn $\neg E$" (sec. 2). I need lose it, though, only in the sense that the information won't be given by my desired credences. Conceivably I might have the information in some other form. One form the information might take is in the double bookkeeping that Arntzenius proposes, having as one's information both one's "epistemic" and one's "prudential" utilities. If, on the other hand, the Swanson report relation R is one-to-one, the needed updating rule will require only the information that standard conditionalization requires.

Arntzenius draws the lesson, "if one's epistemic utilities are linear, then maximizing the expected epistemic utility (by one's current lights) of one's degrees of belief can make it impossible to maximize the expected epistemic utility (by one's current lights) of one's degrees of belief at a future time" (sec. 2). He himself, though, goes on to propose a way out, and it is important to bear in mind two qualifications to what I just quoted. First, we can imagine updating in a way that achieves both these goals if the policy can draw on enough information, as with Arntzenius's own proposal of keeping double books.

Second, some perverse cases differ from the linear one that Arntzenius is treating, in that the Swanson report relation R is injective. For these cases, we don't face this dilemma.

I mostly agree with Arntzenius about his suggested way out, his proposal of keeping two books with two different arrays of credences. An agent who acts as well as believes will need "prudential" credences anyway, to guide her actions in pursuit of new evidence. The Schervish result shows that purely for guidance, the rational agent will want the credences she has. If she also values some form of closeness to truth in her epistemic credences, just for its own sake, she might indeed then wish she kept such double books, with one array of credences to guide her and another to maximize closeness to truth by the standards she embraces. She might wish this, Arntzenius shows, even if she has no other goal than closeness to truth on some specification.

I agree with Arntzenius too that such a wish is ridiculous. First, of course, it will satisfy the believer's preferences only if she cares intrinsically solely about her "epistemic" credences and not about her guiding "prudential" ones. Otherwise, she'll have to find some array of guiding credences that best answer a balance of competing demands: the demand to govern her assessments of expected epistemic utility, and the demand of being truthful in the way she values intrinsically. (Like things would go for wanting credences that will comfort one, enhance one's social dominance, stave off depression and anxiety, and the like. The best thing might be to keep one's epistemically rational credences for purposes of guidance, and have a separate set of cuddly or enlivening credences for these "side" purposes.) Second, if she had the "epistemic credences" she wishes for, they would be idle.

One interesting lesson that Arntzenius draws is worth stressing. He has given, he says, "a purely epistemic argument for updating one's prudential degrees of belief by conditionalization, on the grounds that such updating guarantees cross-time consistency of epistemic utility maximization" (sec. 2). Even if one's goals are purely epistemic, he shows, epistemically rational credences can offer prospectively optimal guidance in achieving those goals. They can do so not only by guiding action in pursuit of new evidence, but by guiding assessments of possible epistemic states for their prospective closeness to truth by some standard. In these senses, we can have a purely epistemic vindication of epistemic rationality.

3. Epistemic Utilities and Coherence

Arntzenius in Section 3 questions the whole notion of epistemic utilities. I should be happy with such questioning: the lesson I drew was a debunking one. Whether or not talk of epistemic utilities makes sense, I argued, no such utilities play any role in explaining epistemic rationality. (I would now admit an exception to this, namely the

roles epistemic utilities played in the last paragraph above.) What might play such a role, I said, is rather a tie to mundane, non-epistemic utilities—to the utility of happiness, wealth, health, or some other such things. I admit I can't myself shake off a residual sense that a pure concern for truth is intelligible and might sometimes be reasonable. Nothing in my debunking, though, required making precise sense of the line I found wanting.

Arntzenius imagines an immobilized robot Hal, and has me asking, "Suppose you just wanted Hal's current degrees of belief to be accurate, what degrees of belief would you give him?" That depends on what I mean by "accurate", he responds—my point exactly. "Gibbard is asking an unclear question." Yes, but as Arntzenius goes on to recognize, I was asking questions like this in order to expose them as unclear. According to Arntzenius, though, I still think the question to be well-defined, though with only person-specific answers. I wouldn't put it that way, and I'm not clear just what such a thought would amount to. My point was that this ill-defined question suggests a whole family of well-defined questions. Tell us just what you mean by "accurate" and you will have indicated a particular question in this family.

Why then have degrees of belief? A big question, this, which I didn't vaunt myself as able to answer. As I think Arntzenius sees, he and I are pretty much in accord on this. "When one's only goal is truth why should one's epistemic state satisfy the axioms of probability?" To this I offered no answer. In the first place, I considered credence just in a single proposition, and so most of those axioms didn't come into play. In the second place, my aim was to refute a certain kind of purely epistemic vindication of standard coherence, and unless some replacement is found, that leaves only the familiar sorts of pragmatic vindications: Dutch book arguments and more comprehensive representation theorem arguments. I may be more optimistic about representation theorems than Arntzenius is, but that's another story, and his expertise on such matters far exceeds mine.

"Why think a rational person must have purely epistemic preferences over all possible belief distributions?" There's no reason—or at least no reason they can't all be zero—unless intrinsic curiosity is itself a requirement of reason. If it is, then the fully rational person is prone act, in some conceivable circumstances, just to find something out, for no further reason. Having learned from Arntzenius of the Hair-Brane theory, I'm curious, and given the opportunity, I might expend resources and effort to garner evidence of its truth or falsehood. Does this require a full set of utilities over my possible states of belief? The story here would be the same as with the rest of decision theory. On the one hand, I can cross bridges when I come to them, and form no preferences until I need them. If, though, I go to an extreme of looking before I might leap, deciding in advance every decision problem that is even conceivable, then consistency may require fully determinate utilities for everything.

If I do have well-defined utilities for everything, can we separate out a purely epistemic component of those utilities? I don't know. My own question was a hypothetical one about a being whose *sole* intrinsic concerns are with her degrees of belief. The being, I supposed, is ideally coherent in her credences. Such a being, I now agree, will still need epistemically rational credences for purposes of guidance. Only epistemically rational credences, after all, will be prospectively optimal, by the being's own lights, as guides in seeking out evidence or assessing the value of possible states of credence. If, though, the being is passive, with nothing she can do but sit back and await new evidence, then thirst for truth as such can't explain her epistemic rationality. Epistemically rational credence can't be explained just as aimed at truth.

4. The Other Puzzle

What, then, of guidance value? The two commentaries focused on the negative thesis of the paper, on the puzzle, if I am right, that aiming at truth as such can't underlie epistemic rationality. The Schervish results lead, though, to another puzzle. Does guidance value somehow underlie the nature of epistemic rationality? I haven't yet seen to the bottom of all this, and I need help.

The main Schervish result is striking: Epistemic rationality is what a fully coherent person will want if she is concerned with her epistemic states solely as guides. Epistemic rationality isn't everything one could want from one's beliefs: one can want comfort, or self-affirmation, or any of a host of other things, and one can want truths just for the sake of having them. Guidance value is just one component of the value that one's beliefs may have. Schervish, though, demonstrates a tight relation between guidance value and epistemic rationality, and it would be strange if the nature of epistemic rationality has nothing to do with this striking relation. But although it is *as if* an epistemically rational person had chosen her credences for the sake of guidance, of course she didn't. She couldn't indeed have conducted a full, rational analysis of prospective guidance values without epistemically rational credences already in place. Exactly what, then, if anything, *is* the bearing of the Schervish findings on the nature of epistemic rationality? That is a second puzzle.