Eric Swanson
ericsw@umich.edu
Note on Gibbard, “Rational Credence and the Value of Truth”
Revision of 10/25/06

Gibbard observes that “With an epistemically rational person, it is as if, by her own lights, she were aiming at truth,” and argues that although aiming at guidance value is sufficient for this kind of “epistemic immodesty,” aiming at truth alone is not. His arguments trade on an analogy between a certain kind of idealized believer and ordinary believers like us: if it is (in certain respects) “as if” we are such idealized believers, then there is good reason to think that we have (certain of) their features. Here I try to undermine Gibbard’s case by showing that for another kind of idealized believer—a kind that is more like us than Gibbard’s believers are—in many cases having the aim of truth alone does suffice for epistemic immodesty. In particular, a believer who ‘aims at truth’ in part by being sensitive to new evidence in the way that is most conducive to the *eventual* accuracy of her beliefs most prefers her actual credences. I don’t think this conclusively shows that our having the aim of truth suffices for epistemic immodesty. But it does make me suspect that Gibbard’s conclusion that guidance value plays a special role in securing our epistemic immodesty is an artefact of his choice of idealization.

1.

Let $g_1(\cdot)$ be a function from a believer’s credence in some proposition S to her value for having that credence if S is true, and let $g_0(\cdot)$ be a function from the believer’s credence in S to her value for having that credence if S is false. Gibbard says that a believer’s valuing “truth and truth alone in her credence in S … seems to consist in satisfying”

Condition $\mathcal{T}$: $g_1(x)$ strictly increases with x increasing, and $g_0(x)$ strictly increases with x decreasing. (5)

Thanks to Frank Arntzenius, Allan Gibbard, Alan Hájek, and Sarah Moss for discussion and helpful comments on earlier drafts.

In many respects Condition $\mathcal{T}$ is not a substantive constraint. Note, for example, that for any positive m and n it is satisfied by the value functions

$$g_1(x) = x^m$$

$$g_0(x) = 1 - x^n$$

The claim that valuing truth alone is compatible with *such* a wide range of pairs of value functions should be controversial. But this is not to say that Condition $\mathcal{T}$ is toothless. Indeed, I think some argument is needed to show that the value functions of a believer who values truth and truth alone must be *strictly* monotonic, as Condition $\mathcal{T}$ demands. Consider for example a believer who, as her known last act, chooses credences that will maximize expected epistemic value by the lights of the value functions

$$\hat{g}_1(x) = \begin{cases} 1 \text{ if } x = 1; \\ 0 \text{ otherwise.} \end{cases}$$

$$\hat{g}_0(x) = \begin{cases} 1 \text{ if } x = 0; \\ 0 \text{ otherwise.} \end{cases}$$

Has such a believer ipso facto ceased to value truth? To be sure, she values correct *guesses* at S's truth value while disvaluing accurate estimates of its truth value, in the sense of Jeffrey 1986 and Joyce 1998. But I find it plausible enough that choosing known-to-be-final credences that are not 'lukewarm' can count as a way of aiming at truth alone.

At any rate, Gibbard thinks that epistemic rationality puts far more substantive constraints on credal value functions. In particular, he thinks that for an epistemically rational agent $g_1(\cdot)$ and $g_0(\cdot)$ must be a **credence eliciting pair**, where this means that a believer with such value functions most prefers to have the credence she actually has (8). I will say that such a believer is **epistemically immodest** with respect to S. Many pairs of value functions satisfy Condition $\mathcal{T}$ without being credence eliciting, and indeed many plausible strengthenings of Condition $\mathcal{T}$ admit non-credence eliciting pairs of value functions.

For example, one might think that a believer who values truth alone in her credences must have symmetric value functions, in the sense that for any $x \in [0, 1]$,

$g_1(x) = g_0(1-x)$.[1] After all, the value of believing S if S is true *just is* the value of disbelieving $\bar{S}$ if $\bar{S}$ is false, and it seems plausible that ways of valuing pure credal accuracy should not be sensitive to the particular proposition that is believed or disbelieved. To motivate this idea in a slightly different way, perhaps "Belief aims at truth" is a special case of the less homey truism that credence aims at accuracy. And a valuation of credal accuracy should not arbitrarily privilege credence in truths or credence in falsehoods by valuing them asymmetrically.

We would then have

CONDITION $\mathcal{T}$, SECOND PASS:

- $g_1(x)$ strictly increases with x increasing, and $g_0(x)$ strictly increases with x decreasing;
- for all $x \in [0, 1]$, $g_1(x) = g_0(1-x)$.[2]

One non-credence eliciting pair of value functions that satisfies SECOND PASS is

$$g_1(x) = x$$

$$g_0(x) = 1 - x$$

As Gibbard notes, a believer with this pair of credal value functions would maximize her expected value by "making [her] beliefs extreme in their certitude" (8) unless her initial credence in S is 0.5. So even SECOND PASS is satisfied by pairs of value functions that are not credence eliciting.

2.

Pairs of credal value functions that make the counterintuitive prescription that we set any credence besides 0.5 to one of the extreme values of 0 and 1 are in some intuitive sense credally pernicious. A believer with such values will, if she can, at a given

[1] See WINKLER 1994 for some discussion of this sort of symmetry.

[2] I also assume henceforth that credal value functions are well-defined and continuous over $[0, 1]$. Strictly speaking I think this assumption does need some argument, but I doubt Gibbard would contest it.

time choose credences that dramatically misrepresent the evidence that she has in fact acquired to that time.

There are several factors that together constitute the credal perniciousness of these particular value functions, however, and it is important to pull them apart. The **report relation** for a pair of credal value functions $g_1(\cdot)$ and $g_0(\cdot)$ is that relation R such that $\alpha R x$ iff, according to $g_1(\cdot)$ and $g_0(\cdot)$, given initial credence α in S, having credence x in S (or reporting credence x in S) maximizes expected value.[3] The report relation for $g_1(x) = x$ and $g_0(x) = (1 - x)$ is:

$$\alpha R \begin{cases} 1 \text{ if } \alpha > 0.5; \\ 0.5 \text{ if } \alpha = 0.5; \\ 0 \text{ if } \alpha < 0.5. \end{cases}$$

R has three properties that encapsulate the credal perniciousness of $g_1(\cdot)$ and $g_0(\cdot)$:

1. Some $\alpha \neq \beta \in [0, 1]$ bear R to the same x. R thus *conflates* prior credences.
2. Some values in $(0, 1)$ bear R to 0, and some bear R to 1. So applying R to a regular credence distribution will sometimes result in an irregular distribution. Even if regularity in one's credences is not a necessary condition for rationality, it is counterintuitive to value irregularity on purely epistemic grounds.
3. Some values in $[0, 1]$ do not bear R to themselves. This is just what it means to have a pair of credal value functions that is not credence eliciting.

The first two properties mentioned above are artefacts of the particular non-credence eliciting value functions we are considering. So it will be helpful to consider pairs that satisfy Second Pass and exhibit only the third property.

Consider for example the following pair of credal value functions, superficially similar to those for the Brier score.

$$g_1(x) = 1 - (1 - x)^3$$

[3] It is important to think in terms of report relations instead of report functions because for some credal value functions distinct credences in S yield maximal expected value given a single initial credence. For example, for $g_1(x) = x^2$ and $g_0(x) = (1 - x)^2$, if $\alpha = 0.5$ we have maximum expected value at both $x = 1$ and $x = 0$.

$g_0(x) = 1 - x^3$

This pair has the report relation plotted in Figure 1: αRx iff $\alpha = \frac{x^2}{2x^2-2x+1}$. For all

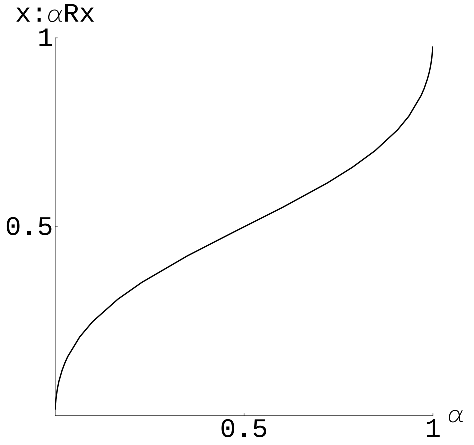

Figure 1: The report relation for $g_1(x) = 1 - (1-x)^3$, $g_0(x) = 1 - x^3$

$\alpha, \beta \in [0, 1]$, $\alpha = \beta$ iff α and β bear R to the same x. Moreover, no values between 0 and 1 bear R to 0 or 1. Nevertheless, this pair of credal value functions is not credence eliciting: for every value of α but 0, 0.5, and 1, R is not reflexive.

Can a believer aim at truth, choose credences partly on the basis of these value functions, and update those very credences as new evidence comes in? If a believer is certain that she will get no more evidence that will interact with her level of credence in S—as it were, if she knows that she is on her deathbed and for whatever reason wants to hedge her bets, somewhat, with her final credences—I think she could choose credences on the basis of this pair of value functions and still count as aiming at the truth. But the circumstances in which a believer can count as aiming at the truth and shift her credences in this way—without making compensating shifts in her updating procedures—are quite rare. A believer who aims at truth alone in her beliefs and thinks that she might get evidence that will interact with her level of credence in S should take every care *not* to let new evidence directly interact with credences that misrepresent her old evidence. Otherwise she would be distorting her total evidence in a way that would undermine her aim to *eventually* estimate the truth in a way that makes the best use of the evidence available to her.

To see this consider a meteorologist trying to decide what value to report as the probability that a storm will pass over a particular island. She is confident that she updates well on the basis of new information, but for prudential reasons she believes that the greater the probability that the storm will pass over the island, the more she should exaggerate that probability in her report. Imagine that she can handle new information that she acquires using either of the following step-by-step strategies:

Applying R at the end

1. She begins to construct an array, writing her initial credences in the first row.
2. When new information comes in, she writes, in row $n + 1$ under the last complete row n, the values that would be the product of her updating on that information if her priors were given by row n.
3. For her forecast she reports the relevant value of the image of the last complete row under a non-credence eliciting report relation R.[4]

Applying R with each update

1. She begins to construct an array, writing her initial credences in the first row.
2. She writes, in row $n + 1$ under the last complete row n, the image of row n under R.
3. When new information comes in, she writes, in row $n + 1$ under the last complete row n, the values that would be the product of her updating on that information if her priors were given by row n, *and then returns to step 2.*
4. For her forecast she reports the relevant value of the last complete row.

Clearly our meteorologist should adopt the first strategy: she should apply R only at the end, so that it affects her report exactly once. Her aim is to estimate the probability that the storm will pass over the island as accurately as she can and then to determine what probability she should report in order to maximize expected value by the lights

[4] In these examples suppose that any $\alpha \in [0, 1]$ bears R to exactly one x. That is, suppose that the report relation can be understood as a report function that is well-defined over $[0, 1]$.

of her value functions. The second strategy has the potential to lead her far astray—for example, it would make her reported value sensitive to *how many times* she had updated, and this is no part of the way she values estimates of truth value. In brief, the fact that she values estimates of the truth value of a proposition disproportionately depending on their proximity to 1 does not entail that she similarly values what amounts to disproportionate *updating*.

For just these reasons, a believer with non-credence eliciting value functions who aims to eventually estimate the truth in a way that makes the best use of her evidence must have doxastic policies that allow her at least to *emulate* the first strategy. But there is a problem here: a believer with such value functions cannot simply *wait* to apply R until the moment of report, as the meteorologist could. A believer with non-credence eliciting value functions applies R to her credences whenever she acts so to maximize expected value. What constraints does this put on her credal value functions?

Perhaps it would be enough that their report relation be injective. Then one R-step 'backward' from the credences the believer arrived at after taking one R-step 'forward' from her initial credences would return her to her initial credences. The believer's updating procedures could then be modified to update not her actual credences, but their image under the inverse relation of R. She could in effect 'undo' her previous choice with each update. We would then have the following constraint on the value functions of a believer who aims at truth, thinks she may acquire more evidence, and wants to put that evidence to fruitful epistemic use:

Condition $\mathcal{T}$, Third Pass:

- $g_1(x)$ strictly increases with x increasing, and $g_0(x)$ strictly decreases with x decreasing;
- for all $x \in [0, 1]$, $g_1(x) = g_0(1 - x)$;
- the report relation for $g_1(\cdot), g_0(\cdot)$ is injective over $[0, 1]$.

But this is just more grist for Gibbard's mill, because even this very strong condition does not rule out non-credence eliciting report relations. For example, the report relation of the 'Brier cubed' score is injective over $[0, 1]$, and the relevant pair of value functions satisfies the other clauses of Third Pass.

3.

There is a problem with this proposal, however, that points the way toward a constraint that permits only credence eliciting value functions. The doxastic lives of the believers that we have chosen to theorize about consist of sequences of updating and acting so as to maximize expected value. Consider some such believer, who has a non-credence eliciting value function and at some time or times

1. Acts, inter alia choosing new credences;
2. Immediately acts again, inter alia again choosing new credences, partly on the basis of the credences she chose in the immediately preceding action.

For such a believer to 'work backward' to credences that aren't distorted by non-credence eliciting choices, modifying updating procedures to compensate for non-credence eliciting choices of credences is not enough: she also needs to ensure that her *actions*—and in particular, her choices of credence—compensate for her choices of credence. Otherwise the believer will choose a credence for a proposition on the basis of a credence which itself may have been chosen to maximize expected value, which itself may have been chosen to maximize expected value, and so on. In virtue of choosing credences in this fashion—one choice immediately following another—such a believer embodies a one-dimensional discrete dynamical system, the evolution rule of which is $R(\cdot)$. A believer who engages in just one choice of credence takes one step in the dynamical system, so that if she starts with a credence in S of $\rho(S)$, she chooses a credence of $R(\rho(S))$ in S. A believer who engages in two immediately consecutive choices of credence takes two steps, so that she ultimately chooses a credence of $R(R(\rho(S)))$, and so on.

Being able to 'work backward' from the output of such a dynamical system requires much more than that the report relation be injective. For example, it is necessary (though obviously not sufficient) that at least one of the following conditions obtains:

1. The believer knows whether the evolution rule has been applied once or twice.
2. Whether the evolution rule has been applied once or twice doesn't make a difference to how the believer should work backward.

The cognitive lives of believers who satisfy the first condition must be quite transparent to them: they must be able to determine, through introspection, how many times they have acted to maximize expected value. We are unlike such believers in a host of ways. We are *less* unlike believers who do not enjoy such introspective transparency, and thus satisfy only the second condition. To be sure, we are unlike them in important respects as well. But we are *closer* to them than we are to believers who can survey their expected value maximizing actions in the ways necessary to satisfy the first condition.

For a believer to satisfy the second condition, her report relation must not conflate distinct credences under iteration. By this I mean that there must not be distinct credences such that one is related to a value by *one* R-step that the other is related to by *two* R-steps. The only report relation with this feature maps each element in $[0, 1]$ to itself and only to itself. And only credence eliciting pairs of credal value functions have this report relation. More formally: The credal value functions $g_1(\cdot)$ and $g_0(\cdot)$ of a believer who thinks she may get new, relevant evidence, and aims at the *eventual* truth—and thus aims to be optimally sensitive to new evidence as it comes in—must satisfy

Condition $\mathcal{T}$, Fourth Pass: The report relation for $g_1(\cdot)$ and $g_0(\cdot)$ is such that for no $x \neq y \in [0, 1]$ is there any z such that xR^2z and yR^1z.[5]

A pair of credal value functions is credence eliciting iff the pair satisfies Fourth Pass.

⇒ Suppose a pair is credence eliciting. Then xRx for any $x \in [0, 1]$, and if xRy then $y = x$. So for any k, xR^kx, and if xR^ky then $y = x$. So for any $x \neq y$ and any k and l, x does not bear the R^k relation to any value that y bears the R^l relation to. In particular, x does not bear the R^2 relation to any value that y bears the R^1 relation to.

⇐ Suppose a pair of credal value functions, $g_1(\cdot)$ and $g_0(\cdot)$, satisfies Fourth Pass. Then for no $x \neq y \in [0, 1]$ is there any z such that xR^2z and yR^1z. $[0, 1]$ is compact, and $g_1(\cdot)$ and $g_0(\cdot)$ are continuous over $[0, 1]$, so by the extreme value theorem every value in $[0, 1]$ bears the R relation to some value in $[0, 1]$. In particular, every value in $[0, 1]$ bears the R relation only to itself. For suppose

[5] aR^nb iff b is accessible from a by an n length sequence of R-steps. So aR^1b iff aRb; aR^2b iff there is some c such that aRc and cRb, etc.

> not: then for some $x \neq y \in [0, 1]$, xR^1y. There is also some $z \in [0, 1]$ such that yR^1z. But then xR^2z, contradicting our initial supposition. So xRx for any $x \in [0, 1]$, and if xRy then $y = x$. So the pair is credence eliciting.

This shows that we are more like believers who (in order to aim at eventual truth) must have credence eliciting value functions than we are like believers who can aim at eventual truth without having credence eliciting value functions.

4.

How should what we learn about hypothetical believers (who always act to maximize expected value, can choose their own credences, and so on) inform our thinking about believers like us? One reason to think about hypothetical believers in general is that they—or at any rate, some of them—help provide tractable and not too misleading models of believers like us. In light of this it would be interesting to see examples of believers whose cognitive lives *demand* that they be modeled using non-credence eliciting value functions: believers for whom it really is "as if" they choose their own credences according to such functions. I am not sure that there are any such believers because, as I have tried to bring out, a believer who has non-credence eliciting value functions and can choose his own credences engages in very odd doxastic behavior over time.

But even without such examples, clarifying the constraints that govern the spaces of various kinds of purely hypothetical believers can point the way toward interesting hypotheses about non-idealized believers. Studying believers that are unlike us can be misleading, however, if we do not correct for artefacts generated by the particular kind of hypothetical believer we choose to focus on. I have argued that for believers who cannot survey the number of times they have acted to maximize expected value, having the aim of eventual truth suffices to ensure that their value functions are credence eliciting. Of course this is compatible with the claim that for *another* kind of believer the aim of eventual truth does not so suffice. But because the believers Gibbard focuses on are in important respects more unlike us than the kind I have discussed, I am not moved to think that it is the aim of maximizing prospective guidance value that secures our epistemic immodesty.

References

Jeffrey, Richard C. 1986. "Probabilism and Induction." *Topoi*, vol. 5: 51–58.

Joyce, James M. 1998. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science*, vol. 65 (4): 575–603.

Winkler, Robert L. 1994. "Evaluating Probabilities: Asymmetric Scoring Rules." *Management Science*, vol. 40 (11): 1395–1405.