THE BADNESS OF BEING CERTAIN OF A FALSEHOOD IS AT LEAST 1/(log 4 – 1) TIMES GREATER THAN THE VALUE OF BEING CERTAIN OF A TRUTH

Alexander R. PRUSS

ABSTRACT: Surprisingly precise results are provided on how much more one should disvalue being wrong than one values being right.

KEYWORDS: credence, epistemic utility, probability

1. Introduction

William James famously thought that epistemic agents differ in how much they comparatively hate error and love truth. Some people "regard the chase for truth as paramount, and the avoidance of error as secondary; or we may, on the other hand, treat the avoidance of error as more imperative, and let truth take its chance."[1] He thought that being a person of the second sort was reasonable. However, it seems that he was wrong. On reasonable assumptions, and bracketing non-epistemic utility considerations, we can show that a rational agent should 'hate' or disvalue being certain of p if p is false at least $1/(\log 4-1)\approx 2.588$ times as much as she 'loves' or values being certain of p if p is true. More generally, if $r\geq 1/2$, one should 'hate' having credence r in p when p is false at least $(2r-1)/(1-2r+\log 4+2\log r)$ times as much as one 'loves' having credence r in p. For instance, you should 'hate' assigning credence 0.95 to a falsehood more than 2.345 times as much as you 'love' assigning credence 0.95 to a truth.

It is surprising that such precise results can be obtained. They will be obtained as a corollary of a necessary condition on proper concave epistemic utility functions.

Normally, epistemic utility functions measure the epistemic value of one's credences given what the truth of the matter is. In this paper, our focus will be on the epistemic utility of one's credence in a *single* proposition p, however, rather than the epistemic utility of one's epistemic state as a whole. This is all that is

[1] William James, "The Will to Believe," *The New World* 5 (1896): 327–347.

needed for the results about love of truth and hatred of error that are announced in the introduction and it simplifies the notation while focusing us on the essentials.

2. Proper epistemic utility functions

Throughout, fix a proposition *p* of interest. We can measure the epistemic utility of a credence *r* in *p* by a pair of functions. $U_T(r)$ is the utility of having credence *r* in *p* should *p* be true. $U_F(r)$ is the utility of having credence *r* in *p* should *p* be false. These functions measure how much one 'loves' or 'hates' being right or wrong about *p*. We shall allow U_T and U_F to take either finite or infinite values at extreme points. Our main interest is in the case where $r \geq 1/2$.

Normally, scoring rule analyses work in terms of measures of *inaccuracy*–the greater the number, the worse. We shall formulate the results in terms of utilities in order to fit with the value-based considerations driving the analysis, and we shall do so in such a way that a familiarity with the scoring-rule literature is not required in the reader. It is worth noting that the above setting is somewhat more general than typical scoring-rule analyses as it allows that the utility-if-true and utility-if-false functions can differ depending on the proposition *p* in question. Our claims are always about a single proposition *p*.

We will now impose some reasonable conditions on U_T and U_F. The first constraint is uncontroversial:

(a) The function U_T is monotonically increasing and the function U_F is monotonically decreasing.

Our next condition is:

(b) The functions U_T and U_F are continuous on the interval [1/2,1], differentiable on its interior (1/2,1) and finite-valued on [1/2,1).

This assumption could be weakened, but it will make the mathematics more convenient.

The following constraint is a weaker version of a fairly standard, though controversial, assumption about scoring rules:[2]

[2] James M. Joyce, "A Non-Pragmatic Vindication of Probabilism," *Philosophy of Science* 65 (1998): 575-603, James M. Joyce, "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial belief," in *Degrees of Belief*, eds. F. Huber and C. Schmidt-Petri (Dordrecht: Springer, 2009), 263-297.

(c) The function U_F is concave on [1/2,1].

A continuous function f is concave on an interval I provided that $f((a+b)/2)\geq(f(a)+f(b))/2$ for all a and b in the domain.[3] Our concavity assumption (c) parallels a standard but controversial convexity assumption on scoring rules (utilities increase with better match between credences and truth, while scores decrease with better match), but it is weaker than that assumption by being restricted to the case where one assigns credence $r\geq1/2$ to a falsehood.

The concavity assumption (c) is in fact quite plausible given the restriction. Intuitively, if p is false, you lose more – or at least no less – by a fixed increase of credence the closer your credence is to 1. Thus, an increase in credence from 0.50 to 0.55 is mildly unfortunate, an increase from 0.55 to 0.60 is no better and very likely worse, an increase from 0.60 to 0.65 is still no better and very likely worse, and so on, until the extremely unfortunate increase in disutility from 0.95 to 1 when you become *certain* of a falsehood. Generalizing this reasoning to all increments implies the concavity of U_F.[4]

Alternately, one might argue like this. Suppose p is false. It intuitively takes a stronger piece of evidence to return one from credence 0.95 to credence 0.90 than to return one from credence 0.90 to credence 0.85, and so on. Therefore, the loss of epistemic utility in moving from 0.90 to 0.95 is greater than in moving from 0.85 to 0.90, because it is harder to return.

Intuitions might be divided on whether U_F is concave on the full interval [0,1]. I am inclined to think it shouldn't be taken to be concave on [0,1/2]. If it were concave, then the gain in utility in a transition from, say, credence 0.25 in a falsehood p to a credence 0.15 in that falsehood would be at least as great as the gain in utility in a transition from 0.50 to 0.40. But the latter seems a more significant transition: it is a move from being on the fence to having a significant inclination to the truth.

The final constraint is the crucial one:

(d) The pair (U_T,U_F) is proper.

[3] In general, if we have no continuity assumption, to define concavity we need to say that $f(\alpha a+(1-\alpha)b)\geq\alpha f(a)+(1-\alpha)f(b)$ for all a and b in its domain and all α€$(0,1)$.

[4] Suppose a and b are in [1/2,1], and suppose for simplicity that $a<b$. Let $\delta=(b-a)/2$ and let $c=(a+b)/2$. Then $c\geq a$ and the generalized claim in the text implies that $-U_F(c+\delta)-(-U_F(c))\geq-U_F(a+\delta)-(-U_F(a))$, because $-U_F(r)$ is the *disutility* of having credence r in the falsehood p. But $c+\delta=b$ and $a+\delta=c$, so $-U_F(b)+U_F(c)\geq-U_F(c)+U_F(a)$, from which it follows that $2U_F(c)\geq U_F(a)+U_F(b)$ and so $U_F((a+b)/2)\geq(U_F(a)+U_F(b))/2$, which implies concavity.

A pair of utility functions is *proper*, roughly speaking, just in case it is never decision-theoretically rational, in respect of the epistemic utility of these functions, to change one's credences without any further evidence.[5]

Before giving a formal characterization of propriety, we can give an example of an *improper* pair of utility functions, and explain why it is improper. It may seem initially very plausible to choose the linear functions $U_T(r)=r-1/2$ and $U_F(r)=1/2-r$. But this has would have untoward consequences. Suppose *p* is the proposition that a toss of a fair six-sided die will *not* yield 6. Obviously, my credence in *p* should be 5/6. But consider the expected epistemic utility of having credence 5/6 in the die toss. I have probability 5/6 of being right and 1/6 of being wrong. My expected epistemic utility, then, is $(5/6)U_T(5/6)+(1/6)U_F(5/6)=2/9$. But what if I just go out on a limb and am *certain* that the toss won't yield 6? My expected epistemic utility, then, is $(5/6)U_T(1)+(1/6)U_F(1)=1/3$. And $1/3>2/9$. More precisely, one can easily check (say, by drawing a graph) that the expected epistemic utility maximizing credence in this case is 1. But it's perverse to switch one's credence from 5/6 to 1 in this case, and any pair of utility functions that recommends it is perverse, or at least *improper*.[6]

Formally, we say that the pair is proper provided that for each r in [0,1], the expected utility function $U(x;r)=rU_T(x)+(1-r)U_F(x)$ has a maximum at $x=r$.

Propriety can also be seen to follow from a continuity assumption on U_T and U_F and two constraints on one's rational method for assigning credences, along with the assumption that there *is* a rational method for assigning credences. We want our rational method for assigning credences to satisfy two plausible criteria. The first is 'precision': the method can potentially return any real-numbered credence value in the interval (0,1). After all, for any rational number, we can easily imagine a lottery situation where that rational number represents the obviously correct credence. The second is 'stable utility maximization': if the method yields some credence value, maximization of epistemic utilities based on that returned credence will not require one to assign some other credence.[7]

[5] For a discussion in the context of scoring rules, see, e.g., Don Fallis, "Attitudes toward Epistemic Risk and the Value of Experiments," *Studia Logica* 86 (2007): 215–246.

[6] Cf. the die example in Joyce, "Accuracy and Coherence," 283.

[7] An anonymous reader suggested that one might want convergence rather than stability. But then the rational method for assigning credences will be to choose the value that is being converged to, rather than the method a single iterative step. And once we choose the value that is being converged to, we will still want stability to apply.

Consider, then, a method *m* of assigning credences that satisfies these two constraints. Given stable utility maximization, $U(x;r)$ must have a maximum at $x=r$ for every r that *m* can return, and given precision, every rational-numbered value in (0,1) must be returnable. To show that propriety follows, we just need to extend this to the endpoints $r=0$ and $r=1$ as well as to irrational values of r. If U_T and U_F are continuous, then $U(x;r)$ is a continuous function of x, and a simple limiting case argument shows that $U(x;r)$ has a maximum at $x=r$ even if $r=0$ or $r=1$ or r is irrational.

Following ideas of Joyce[8] one can also argue for propriety at least in the case of some special *p* by using a special case of Lewis's Principal Principle.[9] Suppose I know for sure that a stochastic process now beginning has a chance *r* of resulting in outcome *A* and a chance 1−*r* of resulting in *B* instead. Let *p* be the proposition that *A* will results. Then by the Principal Principle I should assign credence *r* to *p*. But if the pair is not proper, then assigning credence *r* to *p* in cases like this is not what maximizes objectively expected epistemic utility. Hence, if our credence assignments in such cases are both to match the Principal Principle *and* maximize objectively expected epistemic utility, the utility pairs should be proper.

Standard examples (after transposing from the scoring rule context) of proper pairs are the Brier rule which in our setting will correspond to $U_T(r)=1/4-(1-r)^2$ and $U_F(r)=1/4-r^2$ and the logarithmic rule which in our setting will correspond to $U_T(r)=\log r+\log 2$ and $U_F(r)=\log(1-r)+\log 2$.

Now we have the following simple result, where g' is the derivative of the function g:

Theorem 1: *If U_T and U_F satisfy (b) and (d), then $U_T'(r)=(1-r^{-1})U_F'(r)$ for r in (0,1) and so* $U_T(r)=U_T(1/2)+\int_{1/2}^{r}(1-u^{-1})U_F'(u)\,du$.

[8] Joyce, "Accuracy and Coherence," 279. Alan Hájek ("Arguments for – or against – Probabilism," *British Journal for the Philosophy of Science* 59 (2008): 814) criticizes Joyce's use of the Principal Principle on the grounds that it is not clear that a whole probability assignment could correspond to objective chances, but as an anonymous reader has pointed out this criticism does not apply to the single-case argument I am about to give.

[9] David Lewis, "A Subjectivist's Guide to Objective Chance," in *Studies in Inductive Logic and Probability*, Volume II, ed. Richard C. Jeffrey (Berkeley: University of California Press, 1980), 263–293.

This follows simply from the fact that if $U(x;r)$ is maximized at $x=r$, then the derivative $\frac{dU(x;r)}{dx}$ must vanish at $x=r$, which derivative is equal to $rU_T'(x)+(1-r)U_F'(x)$, so that if this is zero at $x=r$, we must have $rU_T'(r)=(r-1)U_F'(r)$, from which the first result in Theorem 1 follows. The second conclusion in the Theorem follows immediately from the first. Note that the result holds for functions defined on all of [0,1], and not just [1/2,1], if one extends the differentiability and continuity assumptions.

Finally, let us set a neutral point to our epistemic utilities by supposing:

(e) $U_T(1/2)=U_F(1/2)=0$.

This embodies a substantive assumption that the value of credential equipoise does not depend on whether *p* is true or false (the zero-value is a mere convenience for our later discussion – what matters here is that $U_T(1/2)=U_F(1/2)$). One might perhaps question this in the case of some propositions *p*. Perhaps assigning credence 1/2 to a sceptical proposition, such as that I am a brain in a vat, is epistemically worse if the proposition is false than if it is true. This worry *may* involve a confusion between epistemic and non-epistemic utilities. Moreover, the badness of assigning credence 1/2 to a false sceptical proposition may be accounted for by the fact that this forces one to assign non-high credence to many other propositions, and we should not double count when aggregating the utilities over all the propositions one believes. In any case, assigning supposing $U_T(1/2)=U_F(1/2)$ would have to be the way to go if we wanted our utilities not to depend on the particular proposition.

Given (a) and (e), $U_T(r)>0$ and $U_F(r)<0$ for $r>1/2$.

We can now give the Theorem from which the results mentioned in the introduction follow. Suppose $r>1/2$. $U_T(r)$ measures how much, epistemically speaking, one loves having credence *r* when *p* is true, and $-U_F(r)$ measures how much, epistemically speaking, one hates having credence *r* when *p* is false. So we can define the hate-love ratio $HL(r)=\frac{-U_F(r)}{U_T(r)}$ that measures how much more one hates having credence *r* when *p* is false than one loves having credence *r* when *p* is true.

Theorem 2: *Suppose that* U_T *and* U_F *satisfy (a)–(e). Then* $HL(r)\geq\frac{2r-1}{1-2r+\log 4+2\log r}$ *for r>1/2.*

The proof of this theorem is given in the Appendix. In particular, we get that $HL(1)\geq 1/(\log 4-1)$ and that $HL(0.95)>2.345$.

One might be interested to know whether there is any particular pair U_T and U_F that yields precisely the hate-love ratio on the right-hand-side of the inequality. The answer turns out to be affirmative. First, let $U_F(r)=1/2-r$ for r in $[1/2,1]$. Then define $U_T(r)=\int_{1/2}^{r}(1-u^{-1})U_F'(u)\,du$, as would have to be the case for propriety according to Theorem 1. Since $U_F'(u)=-1$, the integral is easy and yields $U_T(r)=\log r+\log 2-r+1/2$ for $r\geq 1/2$. Then, for symmetry, define $U_T(r)=U_F(1-r)=r-1/2$ and $U_F(r)=U_T(1-r)=\log(1-r)+\log 2+r-1/2$ for $r<1/2$.

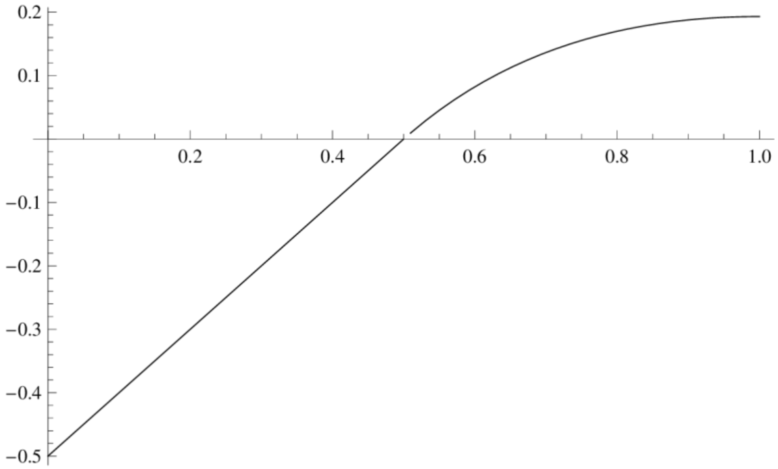

Figure 1: The function $U_T(r)$ that achieves equality in Theorem 2

It is easy to check that the right-hand side of the inequality in Theorem 2 then gives the exact hate-love ratio for this pair of functions. It is easy to see that U_F' is defined everywhere on [0,1] (the point 1/2 is the only place where there could be a problem) and that it is decreasing. Hence U_F is concave on all of [0,1]. In the same way, we can check that U_T is concave.

The remaining thing to do is to check for propriety. Let $U(x;r)=rU_T(x)+(1-r)U_F(x)$. Then $\frac{dU(x;r)}{dx}=r(x^{-1}-1)+r-1$ for $x\geq 1/2$ and

$\frac{dU(x;r)}{dx}=r+(1-r)(1-(1-x)^{-1})$ for $x<1/2$. It is easy to check, considering the cases $r<1/2$ and $r\geq 1/2$ separately, that this derivative is positive for $x<r$ and negative for $x>r$, thereby showing that $U(x;r)$ has a strict maximum at $x=r$. Thus, this piecewise-defined proper pair of utilities achieves the smallest hate-love ratio possible. It might, thus, yield a scoring rule that will be of interest for further investigation.

3. Conclusions

The constraint that one's epistemic utility functions be proper, i.e., that it never make it irrational to stick to one's credences by the lights of these credences, together with a concavity constraint on the $r\geq 1/2$ part of the epistemic utility of believing in a falsehood, is sufficient to determine that these functions need to lopsidedly disfavor believing falsehoods over believing truth.

Plato famously argued in the *Republic* (587b–e) that the true king is 729 times happier than the tyrant. It may seem ridiculous that an exact number of such sort should appear in ethics. Yet, surprisingly, very natural assumptions do occasionally yield various numbers, as in our result that the epistemic disvalue of being certain and wrong is at least $1/(\log 4-1)$ times the epistemic value of being certain and right.[10]

Appendix: Proof of Theorem 2

First we need the following Lemma. It is basically a version of the FKG inequality, but I will need it under slightly different assumptions than those normally used in the FKG inequality, and hence I give a proof from scratch.[11]

Lemma 1: *Suppose f and g are non-negative functions on some interval [a,b], with f monotone non-decreasing and g monotone non-increasing. Then:*

$$\int_a^b f(x)g(x)\,dx \leq \frac{1}{b-a}\int_a^b f(x)\,dx \cdot \int_a^b g(x)\,dx.$$

[10] I am grateful to Lara Buchak, Trent Dougherty, Jonathan Kvanvig and an anonymous reader for relevant discussion and/or comments.

[11] For the standard FKG inequality, see Geoffrey Grimmett, *Percolation* (New York: Springer, 1989), 34. In our setting we might not have the square-integrability assumption.

This lemma says that if f and g are monotone in opposite directions, we won't increase the integral of their product if we replace f with a constant function that has the same average value $(b\text{–}a)^{-1}\int_a^b f(x)\,dx$.

Proof of Lemma 1. For simplicity, assume that a=0 and b=1. The general case follows by rescaling. Use 1_A to denote the indicator function of the set A, i.e., a function that is 1 on A and 0 outside (with a contextually indicated domain). Suppose A is either $[0,\alpha]$ or $[0,\alpha)$ for some α in $(0,1]$. Then:

$$\begin{aligned}\int_0^1 1_A(x)f(x)\,dx &= \int_0^\alpha f(x)\,dx \\ &\le \int_0^\alpha f(x/\alpha)\,dx=\alpha\int_0^1 f(u)\,du \\ &= \int_0^1 f(x)\,dx\cdot\int_0^1 1_A(x)\,dx,\end{aligned} \tag{1}$$

where the inequality followed from the fact that f is monotone non-decreasing and the subsequent equality followed by the change of variables $u=x/\alpha$. The overall inequality also trivially holds if α=0.

Now, let $g_\lambda = \{x\in[0,1] : g(x)>\lambda\}$ be the level set of g. Then for all x:

$$g(x)=\int_0^\infty 1_{g_\lambda}(y)\,dy. \tag{2}$$

Observe also that g_λ is always an interval of the form $[0,\alpha]$ or $[0,\alpha)$ as g is non-increasing, and so (1) applies and yields:

$$\int_0^1 1_{g_\lambda}(x)f(x)\,dx\le\int_0^1 f(x)\,dx\cdot\int_0^1 1_{g_\lambda}(x)\,dx. \tag{3}$$

Since one can reorder integrals of non-negative functions, we can use (2) twice and (3) once to conclude:

$$\int_0^1 f(x)g(x)\,dx = \int_0^1 f(x)\int_0^\infty 1_{g_\lambda}(x)\,d\lambda\,dx$$

$$=\int_0^\infty \int_0^1 1_{g_\Lambda}(x) f(x)\,dx\,d\Lambda$$

$$\leq \int_0^\infty \left(\int_0^1 f(x)\,dx \int_0^1 1_{g_\Lambda}(x)\,dx \right) d\Lambda$$

$$=\int_0^1 f(x)\,dx \cdot \int_0^1 \int_0^\infty 1_{g_\Lambda}(x)\,d\Lambda\,dx$$

$$=\int_0^1 f(x)\,dx \cdot \int_0^1 g(x)\,dx.$$

And in the special case where $a=0$ and $b=1$, that is the desired result. □

Proof of Theorem 2. For brevity, write $D_F(r)=-U_F(r)$ (this is the *disutility* of having credence r in the falsehood p). Then D_F is a non-decreasing monotone function, and it is convex on [1/2,1]. By Theorem 1 and since $U_T(1/2)=0$, we have:

$$U_T(r)=\int_{1/2}^{r} (u^{-1}-1) D_F'(u)\,du.$$

But u^{-1} is a monotone decreasing function while $D_F'(u)$ is monotone non-decreasing since D_F is convex as U_F is concave. By Lemma 1 and since $D_F(1/2)=0$ we have:

$$U_T(r) \leq \frac{1}{r-1/2} \int_{1/2}^{r} D_F'(u)\,du \cdot \int_{1/2}^{r} (u^{-1}-1)\,du$$

$$=\frac{1/2-r+\log 2+\log r}{r-1/2} \cdot \int_{1/2}^{r} D_F'(u)\,du$$

$$=\frac{1/2-r+\log 2+\log r}{r-1/2} \cdot D_F(r).$$

The fraction in front of $D_F(r)$ here must be positive since it is the integral of a function that is positive on [1/2,1). It follows that $D_F(r)/U_T(r) \geq \frac{r-1/2}{1/2-r+\log 2+\log r}$, which is equivalent to the desired result. □