An Epistemic Utility Argument for the Threshold View of Outright Belief

Kevin Dorst

©2014 Kevin Dorst
An epistemic utility argument for the threshold view of outright belief

This paper is in competition for a graduate student stipend for the 2015 Central Division Meeting of the APA. See the attached letter for confirmation.
An epistemic utility argument for the threshold view of outright belief

June 2, 2014

Ever since Joyce (1998) first proposed the idea, there has been a surge of new “epistemic utility” arguments in defense of a variety of Bayesian theses. There are two basic pieces to any such argument (cf. Pettigrew 2013b: §1). One proposes (1) constraints on the class of acceptable “scoring rules” or “epistemic utility functions” for evaluating a given credal state at a given possible world, and (2) a decision rule (e.g. to maximize expected epistemic utility) that translates these evaluations into normative upshots. One then proves that these two assumptions jointly entail some contentious thesis in epistemology. Using this method various authors have now provided new lines of justification for theses such as probabilism (Joyce 1998; Predd et al. 2009; Leitgeb and Pettigrew 2010), conditionalization (Greaves and Wallace 2006), the Principal Principle (Pettigrew 2013a), and a version of objective Bayesianism (Leitgeb and Pettigrew 2010).

One contentious issue in formal epistemology that has not been explored within the epistemic utility framework is the relationship between credences and outright beliefs. In this paper I attempt to fill this gap by putting forward an epistemic utility argument in favor of the threshold view of outright belief. This is a fairly natural view on which there is some threshold $n \in [0,1]$ such that it is rational to believe $p$ iff one’s rational credence in $p$ is greater than $n$. There are various ways one might supplement this picture: perhaps outright belief can be reduced to confidence above a threshold (Foley 2009); perhaps different propositions have different thresholds for belief; and perhaps the relevant threshold can shift depending on the context (Weatherson 2005; Fantl and McGrath 2010). Everything I say will be consistent with these further claims (and their negations); they are all “threshold views” as I am using the term.

Though the threshold view is quite natural, some theorists think that it faces pervasive counterexamples or leads to unpalatable conclusions; thus there are a variety of competing accounts of outright belief. One proposal is that one believes $p$ iff one would be willing to assert it in the context of inquiry; this is inconsistent with the threshold view if one may assert some claims that have very low probability (Kaplan 1996: Ch 4). Another thought is that outright
belief is constitutively tied to reasoning: perhaps to believe \( p \) is to have a dis-
position to treat \( p \) as a true premise in one’s practical and theoretical reasoning
(Frankish 2009; Ross and Schroeder 2014). Or perhaps, unlike credences, out-
right beliefs require causal (as opposed to “merely statistical”) evidence in order

My aim in this paper is not to respond to the objections that lead these
theorists to different views, but rather to propose a new epistemic utility arg-
ument in favor of the threshold view. I show that if we combine (1) certain
natural assumptions about how doxastic states should be “scored” with (2) the
injunction to maximize expected epistemic utility, then it follows that rational
agents will have outright beliefs that conform to the threshold view.¹

Though I am happy to accept this result, the theorists mentioned above may
be inclined to treat my argument as a modus tollens – since they believe that
the threshold view is incorrect, they will infer that there is something unaccept-
able about my assumptions regarding the class of permissible scoring rules. However,
this in itself would be very significant, for – as I will argue in the final section
– my assumptions are entirely analogous to those made by the other epistemic
utility theorists in defense of more traditional Bayesian theses: if one rejects my
results on these grounds, then one must reject theirs as well. Thus my main
conclusion is disjunctive: one must either accept some version of the threshold
view of belief, or reject this fruitful new approach to addressing epistemological
questions.

1 The formalism

Begin with a finite set of possible worlds \( W \). Propositions are simply sets of
worlds, so the set of all propositions is the power set of \( W \): \( P(W) \). Note that
since \( W \) is finite there are also finitely many propositions; namely if \( |W| = k \) then
\( |P(W)| = 2^k \).

We assume that our agent has a probabilistic credence function \( cr(\cdot) \) defined
over the set of propositions \( P(W) \). A belief-state \( B \) is simply a (not necessar-
ily consistent) set of propositions, thought of as the set of those propositions

¹ My results are somewhat similar to some developed by Kenny Easwaran (ms), however
my argument is quite different. Formally, I have a different mathematical base structure and
method of proof, and I generalize the result in a couple important ways. Philosophically, we
are going in opposite directions: Easwaran assumes certain constraints on an agent’s outright
beliefs and goes on to derive a constraint on one’s credences. I, on the other hand, will assume
that one has probabilistic credences and go on to derive a constraint on one’s outright beliefs
– namely, that they conform to the threshold view.

² I make use of this Stalnakerian view of propositions for its simplicity and strength. If one
objects that such a coarse-grained conception of propositions cannot plausibly be the objects
of propositional attitudes, then one might try to generalize my results by adding Hintikka-
style “impossible possible worlds” (1975). Another thought is to try to prove the same results
outside of a possible-worlds framework entirely. But for the purposes of this paper I take a
Williamsonian defense of my approach: developing formal results even in a highly simplified
case can be significant in itself, and it can also open the door to generalizing the strategy
that the agent believes. For example, if \( B = \{p, \neg q\} \) then our agent believes \( p \), disbelieves \( q \), and has no opinion (or suspends judgment) on all remaining propositions. Now, since I am aiming to show that only belief-states that conform to the threshold view are rational, we must get precise on what this requirement amounts to:

**Definition.** A *threshold belief state* \( B \) is one for which there is some \( n \in [0, 1] \) such that for all \( p \in P(W) \): if \( cr(p) > n \) then \( p \in B \), and if \( cr(p) < n \) then \( p \notin B \).\(^3\)

We now need to propose a permissible class of epistemic utility functions for evaluating belief-states at a given world; or, as Jim Joyce puts it, a “criterion of epistemic success” for outright beliefs (1998: 577). He believes that such a criterion is “well-known and uncontroversial”:

**The Norm of Truth:** An epistemically rational agent must strive to hold a system of full beliefs that strikes the best attainable overall balance between the epistemic good of fully believing truths and the epistemic evil of fully believing falsehoods (where fully believing a truth is better than having no opinion about it, and having no opinion about a falsehood is better than fully believing it). (577)

Similarly, Leitgeb and Pettigrew suggest that the fundamental norm governing outright beliefs is simply to “Try to believe truths” (2010: 236). An exceedingly natural way to precisify these thoughts is as follows. First we assign constants \( T, F > 0 \) as, respectively, the degrees to which believing a truth is (epistemically) good and believing a falsehood is bad. Then, since we are working within a finite space of propositions, we can simply count the number of true and false beliefs that an agent has at any particular world, multiply the number of truths by \( T \) and the number of falsehoods by \( F \), and then subtract the latter value from the former.\(^4\) More precisely, let \( \bar{w} \) be the set of propositions that are true at world \( w \). Then we (initially) assume that any acceptable epistemic utility function for outright beliefs will take this form:

\[
U(B, w) = T \cdot |B \cap \bar{w}| - F \cdot |B - \bar{w}|
\]

Note that \( B \cap \bar{w} \) is the set of propositions that are both believed by the agent and true at \( w \), so \( |B \cap \bar{w}| \) is the number of true beliefs in \( B \) at \( w \). Similarly, \( B - \bar{w} \) is the set of propositions that the agent believes but that are not true at \( w \), so \( |B - \bar{w}| \) is the number of false beliefs in \( B \) at \( w \).

In the final section I will return to drawing out the analogies between my assumed utility function and those used by other authors, but for now I will just comment on a couple aspects of this proposal. First, I will assume that the disvalue of believing a falsehood is greater than the value of believing a truth: \( F > T \). Otherwise for every proposition \( p \) and belief-state \( B \) that has no

---

\(^3\)If \( cr(p) = n \) then both belief and non-belief are permissible.

\(^4\)Cf. Easwaran (ms: 5) for an analogous approach to scoring full beliefs.
opinion about \( p \), \( B \cup \{p, \neg p\} \) could be just as good as (or better than) \( B \) itself. Yet surely any plausible utility function will make it so that having inconsistent beliefs about a proposition is worse than having no opinion about it!

Second, one might be puzzled as to why there is no term in my utility function to account for having no opinion on a proposition. This amounts to the assumption that suspending judgment contributes exactly 0 utility to one's overall belief-state. Of course, we plausibly assumed that having no opinion will have an intermediate value between \( T \) and \( \neg F \), but how can I be justified in assuming that this value is precisely 0? Well, suppose that there is a constant \( N \) that we multiply by the number of propositions that an agent has no opinion on in order to contribute to the overall utility of her belief-state. This means that whenever she shifts to a new belief-state that is opinionated with regard to \( p \), she either adds a true belief and gains \( T - N \) utility, or adds a false belief and loses \( F + N \) utility. But then it turns out that there is no need to include \( N \) as a separate term in our equation at all. Instead, we can simply propose new values for \( T \) and \( F \) such that \( T^* = T - N \) and \( F^* = F + N \), and the rankings of belief-states generated by this new function will be equivalent to the original one that had an \( N \)-term in it. (Of course, the numerical value that the utility functions output for a given belief-state may differ, since they are both normalized differently: the former will give \( B = 0 \) a utility of \( N \cdot 2^k \) when \( |W| = k \) whereas the latter function will output a utility of 0. But the change in value between any two belief-states is the same, and so our new function will generate the same rankings as the old one.)

The final piece of background formalism we need is our decision rule for choosing belief-states: namely, maximize expected epistemic utility. The expected utility of a belief-state \( B \) given a credence function \( cr(.) \) and a utility function \( U \) is obtained by multiplying one’s credence in a given world by the utility of the belief-state in that world, and then summing across all such products for the entire space of worlds \( W \). More precisely:

\[
E(B) = \sum_{w \in W} cr(\{w\}) \cdot U(B, w)
\]

(From now on I will abbreviate ‘\( cr(\{w\}) \)’ to ‘\( cr(w) \)’.) Given this, the norm I will rely upon is:

**Maximize expected epistemic utility:** A belief-state \( B \) is (theoretically\(^5\)) rational iff \( E(B) \geq E(B^*) \) for all other belief-states \( B^* \).

Following Greaves and Wallace (2006), I will not provide any independent justification for this norm, other than its intuitive plausibility. If one rejects this piece of my argument, then one is rejecting both orthodoxy and Greaves and Wallace’s powerful justification of conditionalization; I take it that few will find this an attractive alternative.

With these background assumptions and formalism in place, we are now in position to prove my main theorem.

\(^5\)I ignore considerations from practical rationality here and throughout the paper
2 Results

Here is my main result:

**Theorem.** Given a finite space of worlds $W$ and a uniform utility function $U(B, w) = T \cdot |B \cap \bar{w}| - F \cdot |B - \bar{w}|$, a belief-state $B$ maximizes expected utility iff $B$ is a threshold belief-state with $n = \frac{F}{T + F}$.

See the Appendix for proof. The values for $T$ and $F$ determine the relevant threshold, in effect by determining at what level of confidence it is “worth the risk” to believe a given proposition. For instance, if $T = 1$ and $F = 1 + \epsilon$ for an arbitrarily small $\epsilon$, then one ought to believe that $p$ iff $cr(p) > 0.5$. Or if $T = 1$ and $F = 2$, so that the disvalue of a false belief is twice as great as the value of a true belief, then $n = \frac{2}{3}$.

There are a couple worries one might have about this theorem. First, it assumes that the value of $T$ and $F$ are the same for all propositions, which leads to the conclusion that one must have the same threshold $n$ for all propositions. This means that, for instance, if one is equally confident that one’s laundry is dry as that the Higgs-Boson exists, then one should believe the former iff one believes the latter. Yet one might think that this is incorrect: perhaps some sorts of beliefs (unlike those about laundry) are important enough that one must have very high confidence before one can rationally believe them. For these reasons I have elsewhere expanded my results to allow for variable utility functions that allow one to have different thresholds for different propositions. In short, one allows the space of propositions to be partitioned into different domains $S_1...S_m$, and then shows that it maximizes expected epistemic utility to believe a proposition $p \in S_i$ iff $cr(p) > n_i$, where the thresholds $n_i$ are now indexed to the $S_i$.

Second, as it stands my theorem is inconsistent with a pragmatic encroachment view of outright belief on which the threshold for any given proposition can shift depending on the context (usually in response to shifts in practical stakes). However by expanding the class of permissible scoring functions we can likewise expand my results to allow for such views. The basic idea is to add another parameter – a context $c$ – to the utility function, and then index the values of $T$ and $F$ to such contexts. We no longer represent a doxastic state by a belief-state $B$, but rather by a function $\beta(cr, c)$ that takes one’s credence function $cr$ and a context $c$ and outputs a belief-state $B$. An easy corollary of our main theorem is that any belief-state $B$ at any context $c$ maximizes expected utility iff $B$ is a threshold state with $n_c = \frac{F_c}{T_c + F_c}$; thus we get that $\beta$ is rational iff it outputs threshold belief-states for all contexts.

Finally, we could also combine the last two expansions to allow the threshold to vary across both propositions and contexts.

---

For the sake of brevity I do not include this expansion (or the one below) in this paper; I can provide proofs upon request.
3 Conclusion

I have argued that, given certain natural constraints on an epistemic utility function for outright beliefs, a version of the threshold view of belief follows. Though I am happy to accept this result, some may be inclined to treat my argument as a *modus tollens* – since they believe that the threshold view is incorrect, they will infer that there is something unacceptable about my assumptions regarding the class of permissible scoring rules. My two main assumptions were that (1) there are determinate degrees to which true beliefs are good and false beliefs are bad (with these values possibly varying for different propositions and for different contexts), and (2) the overall utility of a belief-state can be determined by summing up the utility gained (lost) for each particular true (false) belief that jointly make up that state. In the current jargon (Pettigrew 2013b; Joyce 2013), (1) and (2) amount to the assumptions that all acceptable utility functions are both **accuracy-centered** and **separable**. The former means that utilities are completely determined by how accurate one’s belief-state is, i.e. how many of one’s beliefs are true and false. The latter means that we can decompose the overall utility of a belief-state into the utilities of the beliefs that make up that state.

Now, I fully admit that these assumptions are not *obviously* correct. One might think that beliefs aim at more than truth – perhaps they aim to conform to the evidence or to constitute knowledge, where these goals are independent of the goal of truth-tracking. Or perhaps one thinks that the epistemic value of a belief-state incorporates holist or coherentist considerations, in which case it cannot be decomposed into the values of the individual beliefs that make it up. These are certainly tenable positions; however they do not undermine the significance of my results. This is because the assumptions of accuracy-centeredness and separability are precisely analogous to those made by the authors mentioned at the outset when they offer their own epistemic utility arguments in favor of traditional Bayesian theses. Though there are some formal differences due to the fact that my argument deals with outright beliefs and theirs deal with credences, all such arguments (with the notable exception of Greaves and Wallace 2006) in effect make (at least) the same two assumptions. For separability, see Joyce (1998: 593); Predd et al. (2009: §3 (Def. 3)); Leitgeb and Pettigrew (2010: 242); and Pettigrew (2013c: 28). For accuracy-centeredness, see Joyce (1998: 591); Predd et al. (2009: §3 (Def. 2)); Leitgeb and Pettigrew (2010: 242); and Pettigrew (2013c: 27).

Given this, I have two points in response to those who wish to treat my argument as a *modus tollens*. First, several of these other authors have defended the assumption of accuracy-centeredness on independent grounds, e.g. by arguing that evidentialist norms can be explained away by deriving them from the goal of accuracy (Pettigrew 2013c; Joyce ms). Second, and more importantly: given the analogies between my assumptions and those of the other authors, one can reject my argument only if one rejects their arguments as well. Thus my main conclusion is disjunctive: one must either accept some version of the threshold view, or reject this fruitful new approach to vindicating traditional (and widely
accepted) Bayesian theses. Either way, we are left with a very significant result.

References


4 Appendix

**Theorem.** Given a finite space of worlds $W$ and a uniform utility function $U(B, w) = T \cdot |B \cap \bar{w}| - F \cdot |B - \bar{w}|$, a belief-state $B$ maximizes expected utility iff $B$ is a threshold belief-state with $n = \frac{F}{T+T_F}$.

**Proof.** Begin by taking a belief-state $B^*$ that is not a threshold state with $n = \frac{F}{T+T_F}$, i.e. there is some proposition $p$ such that either (i) $p \in B^*$ and $cr(p) < \frac{F}{T+T_F}$, or (ii) $p \notin B^*$ and $cr(p) > \frac{F}{T+T_F}$. Now define the ‘switching function’ $S$ that takes a belief-state $B^*$ and a proposition $p$ and outputs a new belief-state that differs from $B^*$ on whether or not it believes $p$. That is, let

$$S(B^*, p) = \begin{cases} B^* - \{p\} & \text{if (i),} \\ B^* \cup \{p\} & \text{if (ii),} \\ B^* & \text{otherwise.} \end{cases}$$

Given this, we start by proving the following:

**Lemma 1.** If $S(B^*, p) \neq B^*$, then $E(S(B^*, p)) > E(B^*)$.

If $S(B^*, p) \neq B^*$, then either (i) $p \in B^*$ and $cr(p) < \frac{F}{T+T_F}$, or (ii) $p \notin B^*$ and $cr(p) > \frac{F}{T+T_F}$. First suppose that (i) obtains. We say that $S(B^*, p) = B^-$, for it omits $p$ while $B^*$ contains it. Now let $w_p$ and $w_{\neg p}$ refer to arbitrary worlds in $W$ at which $p$ is true and at which $p$ is false, respectively. Then we can decompose the expected utility of an arbitrary belief-state $B$ into two components:

$$E(B) = \sum_{w_p} cr(w_p) \cdot U(B, w_p) + \sum_{w_{\neg p}} cr(w_{\neg p}) \cdot U(B, w_{\neg p}).$$

Now note that since $B^-$ has one less true belief than $B^*$ at every $w_p$ world, and has one less false belief than $B^*$ at every $w_{\neg p}$ world, we have:

$$U(B^-, w_p) = U(B^*, w_p) - T$$
and

\[ U(B^-, w_{\neg p}) = U(B^*, w_{\neg p}) + F. \]

Further, since one’s credence in a proposition is simply the sum cross one’s credences in the worlds at which it is true, we have:

\[ cr(p) = \sum_{w_p} cr(w_p) < \frac{F}{T + F}, \]

and since \( cr(\neg p) = 1 - cr(p) \):

\[ cr(\neg p) = \sum_{w_{\neg p}} cr(w_{\neg p}) > 1 - \frac{F}{T + F}. \]

Now compare \( E(B^*) \) with \( E(B^-) \):

\[ E(B^*) = \sum_{w_p} cr(w_p) \cdot U(B^*, w_p) + \sum_{w_{\neg p}} cr(w_{\neg p}) \cdot U(B^*, w_{\neg p}), \]

while

\[ E(B^-) = \sum_{w_p} cr(w_p) \cdot U(B^*, w_p) + \sum_{w_{\neg p}} cr(w_{\neg p}) \cdot U(B^*, w_{\neg p}) \]

\[ = \sum_{w_p} [cr(w_p) \cdot U(B^*, w_p) - T \cdot cr(w_p)] + \sum_{w_{\neg p}} [cr(w_{\neg p}) \cdot U(B^*, w_{\neg p}) + F \cdot cr(w_{\neg p})] \]

\[ = E(B^*) - \sum_{w_p} cr(w_p) + \sum_{w_{\neg p}} F \cdot cr(w_{\neg p}). \]

So the question is whether

\[ \sum_{w_{\neg p}} (F \cdot cr(w_{\neg p})) - \sum_{w_p} T \cdot cr(w_p) \]

is positive or negative. Begin by noting that

\[ TF = TF + F^2 - F^2 \quad \text{(1)} \]
\[ \Rightarrow TF = F(T + F) - F^2 \quad \text{(2)} \]
\[ \Rightarrow \frac{TF}{T + F} = F - \frac{F^2}{T + F} \quad \text{(3)} \]
\[ \Rightarrow \frac{F}{T + F} = F(1 - \frac{F}{T + F}) \quad \text{(4)} \]

Yet since \( cr(p) < \frac{F}{T + F} \) and \( cr(\neg p) > 1 - \frac{F}{T + F} \), we have

\[ F \cdot \sum_{w_{\neg p}} cr(w_{\neg p}) > T \cdot \sum_{w_p} cr(w_p). \]
Thus the value in question is positive, and $E(B^-) = E(S(B^*, p) > E(B^*)$, as desired.

Now turn to option (ii), wherein there is a $p \in P(W)$ such that $p \notin B^*$ and $cr(p) > \frac{F}{T+F}$. Here we say $S(B^*, p) = B^+$, which includes $p$ while $B^*$ omits it. Thus $B^+$ has one more true belief than $B^*$ at $w_p$ worlds and has one more falsehood at $w_{\neg p}$ worlds, so we have

$$U(B^+, w_p) = U(B^*, w_p) + T$$
and
$$U(B^+, w_{\neg p}) = U(B^*, w_{\neg p}) - F.$$

Now from parallel operations to those above we get that

$$E(B^+) = E(B^*) + T \cdot \sum_{w_p} cr(w_p) - F \cdot \sum_{w_{\neg p}} cr(w_{\neg p}).$$

and since in this case $cr(p) > \frac{F}{T+F}$ while $cr(\neg p) < 1 - \frac{F}{T+F}$, by parallel calculations we get that $E(B^+) = E(S(B^*, p)) > E(B^*)$.

So we have now proven our lemma: if $S(B^*, p) \neq B^*$, i.e. $B^*$ is not a threshold belief-state with $n = \frac{F}{T+F}$, then $E(S(B^*, p)) > E(B^*)$.

It is now easy to show that a belief-state $B$ maximizes expected utility iff $B$ is a threshold belief-state with $n = \frac{F}{T+F}$. First, it is clear that all belief-states with threshold $n$ have the same expected utility, for they only differ with regard to propositions such that $cr(p) = \frac{F}{T+F}$, and thesis the point at which believing and failing to believe have the same expected utility. So given that, take any non-threshold belief-state $B^*$. For this to be the case, there must be some propositions $p_1, p_2, \ldots, p_m$ such that either (i) or (ii) holds for each $p_i$. So take $S(B^*, p_1) = B_1, S(B_1, p_2) = B_2, \ldots, S(B_{m-1}, p_m) = B_m$. $B_m$ will be a threshold belief-state and so $E(B_m) = E(B)$. Further, by Lemma 1, $E(B^*) < E(B_1) \ldots < E(B_m) = E(B)$. Since $B^*$ was arbitrary, $B$ has strictly greater expected utility than any non-threshold belief-state with regard to $n = \frac{F}{T+F}$. Thus any such threshold state maximizes expected utility, and any non-threshold state fails to do so.

\[\square\]