Reliability for Degrees of Belief

Jeffrey Dunn

1 Introduction

The concept of reliability is important in epistemology. Most obviously, it is important to *reliabilists* who build accounts of justification or knowledge that make central appeal to the notion of reliable belief production.[1] But even if one does not take the reliabilist line, reliability is still an important concept for epistemological theorizing.

In epistemology, most discussion of reliability has proceeded using *binary models* of belief. These models assume that belief is roughly an all-or-nothing state: either you believe P, you believe its negation, or you withhold belief with respect to P.[2] Within this tradition, different things have been taken to be the *object* of reliability evaluation. Goldman (1979) assesses processes of belief formation (and many have followed him in this); Goldman (1986) assesses entire psychological systems; Greco (1999) assesses the stable dispositions of agents. Despite these differences, there is broad agreement about how to understand reliability. One first specifies a certain set of propositions believed (e.g., the set of propositions in which a process produces belief), and then calculates the truth-ratio for this set: the ratio of true propositions to all propositions. The reliability of the process, the psychological system, or the agent is just this truth-ratio.[3]

[1] For early work on the former, see Goldman (1979); for early work on the latter, see Armstrong (1973).

[2] Sometimes withholding is thought of as the failure to take a doxastic attitude with respect to P; sometimes it is thought of as a third kind of doxastic attitude.

[3] I don't want to minimize the importance of disputes about how exactly to construct these sets. For example, there has been considerable debate about whether or not the truth-ratio for a process should be calculated based on the set of beliefs produced by that process in the actual world, in all possible worlds, or in some special subset of the possible worlds (see, for instance, Goldman (1986), Goldman (1988)). But these disputes take place against a wide amount of general agreement about how to understand reliability. For an alternative view about reliability, which is critical of the appeal to truth-ratios, see Baumann (2009).

Binary models, however, represent only one way of thinking about belief. An alternative view sees belief as coming in degrees. According to these *graded models* there is a whole range of belief strengths between full belief and full disbelief that one can take with respect to a proposition. The central question of this paper is how we should understand reliability when we move from binary models to graded models.

One natural way to understand reliability for graded beliefs is to substitute for truth-ratios the *degree of truth-possession* that a set of graded beliefs has. Roughly, the degree of truth-possession for a graded belief in P of n is the distance that n is from 1 (if P is true), or the distance n is from 0 (if P is false).[4] Though a natural approach, I'll argue that it fails. The reason it fails, however, is interesting. When working with a binary model of belief, a process of belief formation that is reliable both gathers a high ratio of truths to falsehoods and is also highly calibrated with what it is indicating. When we move to graded models, however, these two features come apart. This presents us with *two* distinct ways of understanding reliability: one based on *degree of truth-possession*, the other on *degree of calibration*.[5] I will argue that this calibration-based approach to reliability is preferable and then develop such an approach.

Few authors have investigated this issue directly. Alvin Goldman is an exception, though his work does not paint a clear picture. In some of his work, Goldman has suggested that we should generalize reliability to graded models using degree of truth-possession. For instance, in Goldman (2010) he offers an account of one epistemic system being better than another with respect to binary models of belief. For this account, he appeals to truth-ratios. Just after this he generalizes the account to graded belief "by moving from reliability to the related notion of degree-of-truth-possession, or veritisic value." (p. 194) This strongly suggests that reliability is generalized to graded models by substituting degree of truth-possession for truth-ratios. However, in other places Goldman appears to not understand reliability for graded beliefs as depending on truth-possession. In Goldman (1986), for instance, he floats a calibration-based approach to reliability, though he ultimately rejects it. Very recently, he has indicated some support for the calibration-based approach to reliability for graded belief (Goldman, 2012, p. 26). So, the arguments here are partly critical and partly supportive of the views that Goldman has entertained.

[4]This kind of proposal has roots in the scoring rules of Glenn Brier (1950), Bruno De Finetti (1972), and Leonard Savage (1971).

[5]Very roughly, a process is well-calibrated if it produces beliefs in P to degree n and $n\%$ of the propositions like P are true.

In formulating these arguments, I will simply assume that it makes sense to talk of degrees of belief. I'll also assume that degrees of belief have strengths, which can be represented with numbers between 0 and 1. Thus, degrees of belief can be represented by a function that takes propositions as argument and outputs real numbers in the interval $[0, 1]$. Call such functions *credence functions*. The expression '$c(P) = n$' says that the person in question has a degree of belief of degree n in the proposition P. It will sometimes be useful to distinguish between the fact that someone has a particular degree of belief—or credence—in a proposition, and the credence itself. To mark this distinction, I use the expression '$[c(P) = n]$' to refer to the credal state itself. Although I assume that there are credences, I do not assume that they must be probabilistically coherent,[6] nor do I assume that to have credence in a set of propositions one need have credence in an entire algebra of propositions.

2 Truth-Possession

2.1 Scoring Rules

Suppose that one wants to pursue the truth-possession approach to reliability evaluation for credences. To do this, one first defines a metric for how close a particular token credence is to the truth. Then, one can evaluate a process that produces token credences or a total belief state that contains token credences based on the proximity to truth of those token credences.

The first step requires a scoring rule for credences, and is thus closely related to work on scoring rules for probability estimates.[7] One important distinction in this literature is the distinction between *proper* and *improper* scoring rules. A proper scoring rule is a scoring rule for credences that has the following property: the credence function that has the best expected score from the perspective of any coherent credence function, c, is c itself.[8]

[6] Although perhaps to be *rational* they must. A set of credences is probabilistically coherent just in case they satisfy the standard Kolmogorov axioms:

1. For all P, $c(P) \geq 0$,
2. For all logical truths, $\top$, $c(\top) = 1$, and
3. $c(P \vee Q) = c(P) + c(Q)$ for all P, Q such that $(P \wedge Q)$ is contradictory.

[7] The literature here is vast. For early work, see Brier (1950), De Finetti (1972), and Savage (1971). More recent philosophical work on this topic has been done by Joyce (1998, 2009), Greaves & Wallace (2006), Gibbard (2008), and Leitgeb & Pettigrew (2010a,b).

[8] See Seidenfeld (1985, p. 285).

Let $\mathcal{P}$ be the indicator function for proposition P (i.e., the function that takes value 1 if P is true and 0 if P is false). An example of an improper scoring rule is the Linear Score:

Linear Score: $[c(P) = n]$ is given a score of $|\mathcal{P} - n|$.

An example of a proper scoring rule is the Brier Score[9] (Brier, 1950):

Brier Score: $[c(P) = n]$ is given a score of $(\mathcal{P} - n)^2$.

Notice that the closer one's degree of belief is to the truth, the lower the score from both these functions. Thus, lower scores are better.

To evaluate a process or a credal state for reliability, one then makes use of one's preferred score. For instance, an account of the reliability of processes using the Linear Score would look like this:

Linear Reliability: The average Linear Score for the degrees of beliefs produced by process ρ is the reliability of process ρ.

However, if one is pursuing a truth-possession approach to reliability, there is good reason to reject accounts like Linear Reliability and choose a proper scoring rule instead.[10] To see why, suppose that one adopts Linear Reliability and imagine an agent who is interested in ensuring that her processes of belief formation are as reliable as possible. Suppose that process ρ consistently issues credences of 0.7 in a set of propositions. The agent thus thinks that each proposition in this set is 0.7 likely to be true. Since the Linear Score is improper, the expected value she assigns to having those credences is less than the expected value of having a slightly higher credences in those propositions. She will thus think that her process of belief formation would be more reliable were it to assign slightly higher credence to the propositions. This line of reasoning will continue until the process produces credences of 1 in the set of propositions. This scoring rule, then, forces an agent to regard any process that does not assign propositions extreme values as less reliable than it could be.[11]

[9] Seidenfeld (1985, pp. 285-6) calls this the *Quadratic Loss Function*. It is also very similar to de Finetti's S score (De Finetti, 1972, p. 30).

[10] The truth-possession accounts that Goldman has defended (Goldman & Shaked, 1991; Goldman, 1999, 2010), whether one reads them as analyzing *reliability* or not, all appeal to the Linear Score. The reason given here for not using a Linear Score in an analysis of reliability applies equally to Goldman's use of it.

[11] Michael DePaul (2004) criticizes Goldman (1999) for construing the epistemic good as truth-possession. His strongest criticism relies on the fact that Goldman's proposal requires agents to move their degrees of belief to extreme values in something like the way

A much more promising approach is to use a proper score, such as the Brier Score.[12] If we take this line and continue to evaluate processes for reliability, this yields:

Brier Reliability: The average Brier Score for credences produced by a process is the reliability of the process.[13]

Moving to a proper scoring rule gets one out of the problem above. If your credence in P is n, then you minimize your expected Brier Score by having n as your credence. Thus, agents self-consciously pursuing reliability need not be compelled to adopt processes that assign only 0s or 1s to propositions.

2.2 Problems for Truth-Possession Reliability

Moving to a proper scoring rule (such as the Brier Score) fixes an obvious problem for some accounts of reliability based on truth-possession. Nevertheless, I'll argue that there is a deeper problem that besets all accounts of reliability based on truth-possession, whether the account uses a proper score or not.

The basic problem is this: our notion of reliability should allow that there can be highly reliable processes that yield mid-level credences (i.e., credences not near 0 or 1). More specifically, consider a process that produces credences of the form $[c(P) = 0.6]$ ('Process 0.6') and one that produces credences of the form $[c(P) = 0.9]$ ('Process 0.9'). The claim is that whatever level of reliability Process 0.9 can reach, there should be situations

illustrated here (DePaul calls this 'epistemic swashbuckling'). However, DePaul doesn't attribute the problem to the scoring rule being improper, nor does he consider the modification of using a *proper* scoring rule to get around this problem.

[12]It is worth noting that there are many proper scoring rules (Savage, 1971; Seidenfeld, 1985; Gibbard, 2008). So, although the Brier Score offers one way of pursuing the truth-possession approach to reliability, it is not the only way. Joyce (2009) offers some arguments on behalf of the Brier Score being a privileged choice, but I'll leave this issue to the side, focusing instead on a complication that arises for any approach to reliability evaluation that appeals to truth-possession.

[13]Note that this description does not fully determine an evaluative scheme. In particular, one could take the average Brier Score for a set of propositions by averaging the Brier Score of all the *atomic propositions*, or by averaging the Brier Score of all the elements/worlds of the probability space, or indeed by averaging the Brier Score of some other way of cutting up the probability space. How one decides to compute these average scores can make a difference to the overall verdict. Throughout the body of the paper, I will assume that there is some non-arbitrary way of figuring out what the atomic propositions are and that scores are assigned to the atomic propositions produced by a process. Since I will argue that there is a problem with truth-possession proposals *independent* of this issue, I don't consider it any further.

where Process 0.6 can reach that same level of reliability. However, Brier Reliability does not allow this.

That a concept of reliability should have this feature can be seen by looking at a scenario that makes use of a binary model of belief. Suppose there is a process that produces beliefs in disjunctive propositions about the weather. It issues beliefs in propositions of the form $P \vee Q$ where $P =$ *It will be cloudy today*; $Q =$ *It will be over 70 degrees today.* Such a process could be very reliable in that when it produces a belief in a proposition of the form $P \vee Q$, $P \vee Q$ is normally true. Consider a different process that produces beliefs in propositions that are more informative. It issues beliefs in propositions of the form P. This process gives us more information about the world with respect to P than does the process that produces beliefs in the disjunctive proposition. Nevertheless for any level of reliability that the second process reaches, the first one is able to reach that level of reliability, too. Similarly, Process 0.6 is less informative about P than is Process 0.9. Nevertheless, Process 0.6 should still be able to reach whatever level of reliability Process 0.9 can.

To see that Brier Reliability cannot deliver this, note first that if the set of propositions the processes issue degrees of belief about, $\mathbf{P}$, contains only true propositions, then each process will be getting its best possible score. Process 0.6 gets a score of 0.16 and Process 0.9 gets a score of 0.01 (recall that lower scores are better). This, on its own, is not objectionable. Process 0.9 does seem to be doing better here. Process 0.6 will do worse overall but better than Process 0.9 when there are sufficient number of propositions in $\mathbf{P}$ that are false. This is because although Process 0.6 doesn't get as large a reward when some $P \in \mathbf{P}$ is true as does Process 0.9, it gets a bigger reward than Process 0.9 when some such P is false. Consider, then, a case where 60% of the propositions in $\mathbf{P}$ are true and 40% false. In this case, Process 0.6 gets a score of 0.24 and Process 0.9 gets a score of 0.33. So, Process 0.6 is outperforming Process 0.9 here. It turns out that in this case, Process 0.6 is outperforming *any* process that issues uniform credences to the propositions in $\mathbf{P}$.[14] The problem, however, comes when we compare the score that Process 0.6 gets when it is outperforming all other processes compared to the score that Process 0.9 gets when it is outperforming all other processes. Process 0.9 does this when 90% of the propositions in $\mathbf{P}$ are true. In this case, Process 0.9 gets a score of 0.09. So, according to Brier Reliability, when Process 0.6 is at its best, it is not as reliable as Process 0.9,

[14] In general, Process n outperforms Process k ($k \neq n$) whenever the proportion of true propositions in $\mathbf{P}$ is n.

when it is at its best. Further, as we saw, the best Brier Score for Process 0.9 is better than the best for Process 0.6. But that means that Brier Reliability cannot deliver the desiderata outlined in the previous paragraph: according to Brier Reliability, processes that produce mid-level credences cannot be as reliable as those that produce high-level credences.

There is a second, and related, reason to think that Brier Reliability is mistaken. This reason is closely related to the role that reliability is supposed to play for reliabilism about justification. Reliabilism about justification says that the reliability of a process determines the justificatory status of a belief produced by that process.[15] If, then, processes that produce mid-level credences cannot be produced by highly reliable processes, then such credences could not be highly justified. This, however, is wrong. One would think that there are certain situations where a credence of, say, 0.6 in a proposition is highly justified. The following example brings out this point.

Consider Process a, which delivers credences about propositions in a certain set, $\mathbf{A}$. Suppose that a works by picking up on a feature of the situation, F_A, which is correlated with the truth of the propositions in this class. That is, whenever a detects F_A a produces the belief $[c(A) = 0.95]$. Suppose that when feature F_A is present, the corresponding A is true 95% of the time. Now, consider Process b. It works in a very similar way, but about propositions in a different class, $\mathbf{B}$. Whenever b detects F_B b produces the belief $[c(B) = 0.7]$. Suppose that when feature F_B is present, the corresponding B is true 70% of the time.

For example, Process a might produce beliefs about blue jays. If the process registers a certain blue patch in a tree, it produces a credence of 0.95 that there is a blue jay in the tree. Since there aren't many other blue things in trees, 95% of the time a blue patch is detected, there really is a blue jay. Process b, on the other hand, might produce beliefs about cardinals. If the process registers a certain red patch in a tree, it produces a credence of 0.7 that there is a cardinal in the tree. However, because there are other red things in trees—berries, for instance—the red patch is not as highly correlated with the presence of cardinals. Hence, the lower credence assigned to the proposition that there is a cardinal in the tree.

Finally, consider a modified version of Process a. This modified process is exactly the same as a except that it assigns to the propositions in $\mathbf{A}$ a credence of 0.6. That is, when it detects a blue patch in the tree, it produces

[15] I'll assume that such a view says that the more reliable a process that produces a belief, the more justified it is. This is a plausible thing for reliabilism about justification to say, but it is not universally endorsed. Though it makes the argument easier to present, it is ultimately inessential to it.

a credence of 0.6 that there is a blue jay. It does this despite the fact that when there is a blue patch detected, a blue jay is present 95% of the time. Call this Process a^*.

I think is is clear that a (the blue jay process) and b (the cardinal process) produce credences that are equally justified. That is, assigning a credence of 0.95 to there being a blue jay in virtue of Process a is just as justified as assigning a credence of 0.7 to there being a cardinal in virtue of Process b. But perhaps some will disagree. It seems undeniable, however, that credences produced by a^* are less justified than credences produced by b. That is, assigning a credence of 0.6 to there being a blue jay, in virtue of Process a^* is less justified than assigning a credence of 0.7 to there being a cardinal in virtue of Process b.

However, Brier Reliability when paired with reliabilism about justification cannot deliver these results. The scores for these processes are:[16]

Process a: 0.0475

Process b: 0.21

Process a^*: 0.17

But then reliabilism about justification says that the credences produced by a^* are more justified than those produced by b.

This argument takes for granted (1) reliabilism about justification and (2) Brier Reliability, to derive what I claim is an absurd result that the credences produced by a^* are more justified than those produced by b. There are two main responses open to someone who wants to maintain Brier Reliability. One could argue that the alleged absurd claim is not so absurd. I'll address this objection in a moment. For now, focus on the other main line of response: this argument shows not that Brier Reliability is mistaken, but that reliabilism about justification is. Obviously, this sort of response won't be attractive to reliabilists, who want to use reliability to assess the justificatory status of beliefs. But I think the argument shows that there is something wrong with Brier Reliability independent of one's commitment to reliabilism about justification. The reason: this would be a too-easy refutation of reliabilism about justification. Reliabilism about justification certainly faces challenges. But these challenges take two main forms. One kind of challenge alleges that there is no non-arbitrary way to specify the

[16] Given how a works, we know that 95% of the propositions in $\mathbf{A}$ will be true. We can thus calculate the average Brier Score for Process a: $0.95 \times (1-0.95)^2 + 0.05 \times (0-0.95)^2 = 0.0475$. Similar calculations yield the scores for the other processes.

justificatory status of a belief because there is no non-arbitrary way to specify the relevant process that produced it. This is the generality problem. The other kind of challenge alleges that reliabilism yields unintuitive verdicts about cases where a process has reliability properties unbeknownst to the agent who has formed a belief. These are worries about externalism. But the scenario here involves none of this. We can stipulate that the agent with the processes understands perfectly well how they work. Further, we can stipulate that in this case, it is clear that these are the relevant process types. So, this case doesn't have any of the usual features that generate worries about reliabilism. Further, given that it is not at all clear how to understand reliability for graded beliefs, but that there are many who are attracted to reliabilism about justification, it is premature to place the blame on the latter view, rather than on the particular metric used to calculate reliability.

The other line of response to the argument is to maintain that it is *not* absurd to hold that the credences produced by a^* are more justified than those produced by b. In response, note that given how b works and the features of the environment it responds to, there is no better credence that it could produce than $[c(B) = 0.7]$. That is, it would not be epistemically better if the process produced credence in B to some different degree or just issued no credence whatsoever. But now consider Process a^*. Given how a^* works, there is a better credence that could be produced than $[c(A) = 0.6]$. It would be better if the process produced $[c(A) = 0.9]$. This, it seems, is a good reason to think that the credence produced by b cannot be *less* justified than the credence produced by a^*.[17]

One objection to this response is to agree that b is, in some sense, doing the best it can, but maintain that this doesn't show that the credences it produces are more justified or in any way better than those produced by a^*. Further, the objection goes, this is something with which reliabilists should be familiar. There are bad situations where none of an agent's processes are

[17]This argument is analogous to Stewart Cohen's (Cohen, 1984) "new evil demon problem". Internalists about justification maintain that justification supervenes only on internal features of agents. Externalists about justification deny this. In Cohen's scenario you are to imagine an internal twin of yourself who is living in a demon world. Just like you, your twin believes that he has a hand. But your twin's belief is not produced by a reliable process since the demon is constantly deceiving him. This puts pressure on the externalist to admit that an unreliable process can produce justified beliefs. Why? Because your twin seems to see his hand and in virtue of that believes he has a hand. There seems to be no other belief that your twin could form that would be more justified. Thus, his unreliably formed belief is justified. Similarly, I maintain that since Process b is making the best response it can, the credence produced is justified.

very reliable; in such a situation the agent has few if any justified beliefs. Such an objection, however, misses the key way in which credences are different than binary beliefs. When it comes to binary beliefs, there is only one way in which processes can be differentiated from each other: how correlated the beliefs they produce are with the truth. For credences there is an extra degree of freedom. Processes can be distinguished not only in terms of their correlation with the truth, but also by what level of credence is produced by that process. Given this extra degree of freedom, it is implausible to maintain that in situations of lower correlation, there are simply no justified credences to be had.

In summary, then, I think there are two good reasons to think that Brier Reliability is mistaken. The first reason is that it treats processes that produce mid-level credences differently than those processes that produce high-level credences. The second reason, which is related to the first, is that Brier Reliability when paired with reliabilism about justification issues incorrect verdicts about cases. But this still leaves two important questions. First, why does Brier Reliability go wrong, given that it is simply a generalization of reliability evaluation in binary models? Second, is there some easy technical fix to Brier Reliability just as there was with Linear Reliability? I'll address these questions in turn.

The answer to the first question has already been suggested. Brier Reliability goes wrong because two important features, (1) gathering truth and (2) calibration to the truth-ratios, go together when we are considering binary beliefs and yet come apart when we have graded beliefs. Suppose, for instance, that processes a and b had simply produced full beliefs in their respective propositions. Process a would then, on average, get the agent more true beliefs than b. But notice that a would also be making the better calibrated response to the truth-ratio of propositions in $\mathbf{A}$. It would be better calibrated in the sense that it would lead to acceptance of a set of propositions 95% of which are true, in contrast to b, which would lead to acceptance of a set of propositions only 70% of which are true. It is this latter notion, I argue, that reliability evaluation should be tracking, and it just happens to go along with closeness to the truth when we are working with binary models. When we move to graded models, however, these two features come apart. Brier Reliability aims at generalizing the notion of closeness to the truth and so misses out on what reliability evaluation should be tracking.[18]

Finally, is there a simple technical fix to Brier Reliability to get around

[18]This point connects up in interesting ways with recent work on epistemic value. In his contribution to a recent book, Duncan Pritchard (Pritchard *et al.*, 2010) considers (without endorsing) the following view:

this problem? There is not. *Any* scoring rule for credences that is guided by the idea that possessing more truth is better will run into this problem. This is because if more truth is better, then, like the Brier Score, the rule will give a score to $[c(P) = n]$ that decreases monotonically as n increases (so long as P is true). Though there are many proper scoring rules, *all* will satisfy that requirement, and so *all* versions of this approach to reliability evaluation will run into the objection given above. They will be biased in favor of those processes that garner more truth at the expense of being highly calibrated.[19] Degree of truth-possession may be valuable for certain evaluative purposes, but not for reliability-based evaluation.

3 Calibration

The overarching concern with truth-possession makes the truth-possession approach blind to an important way that a process can be appropriately responsive to its environment. For this reason, degree of truth-possession does not provide an appropriate metric for reliability.

There is, however, an alternative. This alternative method is suggested by Frank Ramsey:

> ... given a habit of a certain form, we can praise or blame it accordingly as the degree of belief it produces is near or far from the actual proportion in which the habit leads to truth. We can then praise or blame opinions derivatively from our praise or blame of the habits that produce them. (Ramsey, 1931, p. 196)

Epistemic Value T-Monism: True belief is the sole fundamental epistemic good.

As stated, Epistemic Value T-Monism doesn't say anything about degrees of belief. But it is not implausible to extend the view as follows:

Epistemic Value T-Monism (degrees): Truth-possession is the sole fundamental epistemic good, the more the better.

Notice that if you are committed to Epistemic Value T-Monism (degrees), then you have to say that from the perspective of epistemic value, Process a^* is better than Process b. To the extent that one is convinced by my arguments here, one has reason to maintain that Process a^* is less reliable than Process b. This either shows that greater reliability need not go together with greater epistemic value, or that Epistemic Value T-Monism (degrees) is mistaken. The latter option, in turn, puts pressure on one to reject Epistemic Value T-Monism. In section 4.2 I consider how one might resolve this.

[19] See Blattenberger & Lad (1985) for a graphical representation of the relationship between calibration and the Brier Score, which demonstrates how one can trade-off calibration for an increased Brier Score.

The idea is that instead of looking to degree of truth-possession, we look at how well calibrated the process is. Roughly, calibration will be measured by comparing the credence value assigned to a proposition by a process to the proportion of true propositions that the process issues credences in. Notice that this will allow that processes that produce mid-level credences can still be just as reliable as those that produce high-level credences.[20]

3.1 Calibration, Refinement, and Proper Scoring Rules

As noted above, the Brier Score is a proper scoring rule. There has been interesting work done on proper scoring rules and their connection with measures of calibration.

Let $\mathbf{P}$ be a finite set of N propositions to which process ρ has assigned credence. Let b_i be the credence assigned to $P_i \in \mathbf{P}$, and let $\mathcal{P}_i$ be the indicator function for P_i. Then, the Brier Score for ρ is:

$$\textbf{Brier Score: } BS(\rho) = 1/N \sum_i (\mathcal{P}_i - b_i)^2$$

Work by Sanders (1963) and Murphy (1972, 1973) shows that this can be decomposed as the sum of two quantities: the *Calibration Score* and the *Refinement Score*.[21] Let $\{\mathbf{P}_j\}$ be a partitioning of $\mathbf{P}$ into disjoint subsets such that $\mathbf{P}_j = \{P \in \mathbf{P} : c(P) = b_j\}$ where each element of $\mathbf{P}$ has a credence in $\{b_1, b_2, \ldots, b_j\}$. Let n_j be the cardinality of $\mathbf{P}_j$ and (as before) N be the cardinality of $\mathbf{P}$. Finally, let r_j be the proportion of $P \in \mathbf{P}_j$ that are true. We can then define these quantities:

$$\textbf{Calibration Score: } C(\rho) = \sum_j (n_j/N)(r_j - b_j)^2$$

[20] As I discuss shortly, this basic idea is considered, though rejected, in Goldman (1986). The introduction to Goldman (2012), however, suggests that Goldman is now sympathetic to such a view. Bas van Fraassen (1983) considers the notion of calibration for credences, arguing that when a credence is calibrated with the frequencies, then it is *vindicated*. This is similar to how an all-or-nothing belief is vindicated when it is true. Alan Hájek (unpublished) also proposes that it is good for credences to be calibrated, but rather than calibrated to the frequencies he suggests they should be calibrated to the objective chances. Neither van Fraassen or Hájek, however, argue that calibration is related to reliability. Instead, both of them think of calibration as replacing degree of truth-possession as the overarching epistemic goal. Marc Lange (1999) also investigates the notion of calibration, however he focuses on agents who *believe* they are calibrated not those who actually are.

[21] See also DeGroot & Fienberg (1983) and Blattenberger & Lad (1985). For a philosophical treatment of this, see Joyce (2009). Schmitt (2000, pp. 265-8) has an informal discussion of some of this material specifically related to Goldman's social epistemology.

$$\textbf{Refinement Score: } R(\rho) = \sum_j (n_j/N)(r_j)(1 - r_j)$$

From this it follows that $BS(\rho) = C(\rho) + R(\rho)$.[22]

In English, this tell us that the Brier Score can be thought of as having two components. The calibration component tells us how closely the credences produced by a process match the frequency of truth among the propositions the process issues verdicts about. It is *this* quantity that I will argue best captures the reliability of a process. The second component of the Brier Score is refinement. It tells us how homogenous the truth-values are of the propositions about which the process issues verdicts. A refinement score of 0 is achieved if either all the P_i are true or all the P_i are false. The worst refinement score, 0.25, is achieved when half the P_i are true and half are false. This means that if a certain process, ρ, happens to be aimed at issuing credences in propositions which are not homogenous with respect to truth-values, then ρ will be penalized for this according to the Brier Score, even if the process is highly calibrated.

3.2 Early Goldman

Goldman (1986) proposes a system of reliability evaluation for credences based on the idea of calibration. In that work, Goldman is centrally concerned with defending a version of reliabilism about justification for binary belief. The view he defends calculates the reliability of total epistemic systems (rather than processes) and does so using truth-ratios.[23] When it comes to graded beliefs, Goldman offers something similar, utilizing the notion of a *well-calibrated belief set*. According to Goldman, a graded belief set is well calibrated iff for each number d, the truth-ratio of propositions assigned a credence of d is approximately d. This is just the Calibration

[22] I focus on the Brier Score. As is well-known, there are other proper scoring rules. In general, any proper scoring rule has a decomposition into a calibration component and a refinement component as illustrated in the text for the Brier Score (DeGroot & Fienberg, 1983, pp. 19–21). I focus in the text on the Brier Score, but I offer no arguments here for its advantages over other proper scoring rules. The primary claim I wish to defend is that calibration rather than truth-possession is a good measure of reliability. This point is independent of the choice of scoring rule. It's worth noting that all proper scores will agree on *perfect* reliability.

[23] ARI is the principle that expresses this. Goldman writes: "**ARI** A J-rule system R is right if and only if R permits certain (basic) psychological processes, and the instantiation of these processes would result in a truth-ratio of beliefs that meets some specified high threshold." (Goldman, 1986, p. 106).

Score for the graded belief set. Goldman then proposes substituting calibration for truth-ratio in evaluating graded beliefs for reliability (Goldman, 1986, p. 114).[24]

Just after introducing the idea, however, Goldman claims that this proposal runs into a problem. Suppose it is true of an agent, that out of all the propositions she has an opinion about, 70% are true. If so, then her credence function achieves perfect calibration by setting her credence equal to 0.7 for all propositions about which she has an opinion. But to say that such a credence function is perfectly reliable, Goldman thinks, is clearly wrong.[25]

One response to this problem is to reject the entire calibration approach to reliability evaluation and move to a truth-possession approach. But another response is to try to fine-tune the calibration approach. Since I have argued that the truth-possession approach is flawed, this is a promising option. I'll first present this fine-tuned response, and then explain how it deals with Goldman's worry, the problem cases that I've argued should lead us to reject Brier Reliability, and other related concerns.

3.3 Calibration Reliability

The view I defend identifies the reliability of a process with the calibration score that it receives. As a reminder, this is given by:

$$\textbf{Calibration Score: } C(\rho) = \sum_j (n_j/N)(r_j - b_j)^2$$

This formal statement of the Calibration Score, however, does not tell us anything about how to construct the set of propositions that ρ has produced credence in and that are being evaluated for their calibration. To see that there is an issue here, consider how we might state the view non-formally:

Calibration Reliability (draft): To determine the reliability of process ρ:

1. Construct a set of propositions that contains all the propositions that ρ has assigned credence to.
2. Partition this set of propositions according to the credence assigned to them.

[24] Since Goldman (1986) only defines perfect calibration, he doesn't take a stand on how to measure the distance from perfect calibration, and so doesn't take a stand on the precise form of the scoring rule.

[25] Seidenfeld (1985) notes this problem with calibration, too, though not specifically with respect to Goldman.

3. Let the score for each partition be the squared difference between the credence and the truth-ratio of propositions in that partition.
4. The weighted average of the scores for each partition is the reliability of ρ.

The important issue concerns step 1. As stated, the view says that to assess the reliability of a process, ρ, we look only at the propositions that ρ has actually assigned credence to. But on reflection, that can't be right. Consider how we evaluate a process for reliability in a binary model. We look at the actual beliefs formed by that process, but also take into consideration counterfactual beliefs that could have been formed. We do this because we recognize that a process of belief formation might accidentally issue beliefs in propositions most of which are true.

For example, suppose that there is a crazy professor who lies to the first, third, fifth, etc. student who asks him a question on a day, and tells the truth to the second, fourth, sixth, etc. student. Suppose Jane routinely ask the professor questions and we want to assess her process of belief-formation for reliability. Even if, as luck would have it, she's always been an even-numbered visitor to the professor, we do not evaluate her process of belief formation as highly reliable. There is, perhaps, a derivative notion of reliability where we can evaluate just the beliefs she's actually formed for how reliable they are. But this is not a philosophically interesting notion of reliability, since it is unable to distinguish lucky accidents from true reliability. Of course, this is still consistent with the fact that one of our best guides to the reliability of a process is its *actual* performance. But in assessing reliability we care about more than just actual track record.

So, just how far from the actual world do we let things vary? That's a difficult question in general, but there are several specific things to say. First, it is the external environment we vary not the process itself. For instance, to evaluate Jane's process, we don't consider what would happen if Jane asked fellow students her questions instead. We do, however, want to vary the specific times that Jane goes to ask her professor questions. Second, we want to hold fixed non-accidental patterns in the world, but let those that are accidental vary. This is a vague idea, but there are clear cases. Laws of nature, for instance, are non-accidental patterns that are held fixed. That everyone in a coffee shop is an only child is an accidental pattern (unless, for example, there's some sort of special meeting going on). In the crazy professor case, that the professor alternates true and false answers to questions is non-accidental (that's what makes him crazy), but the specific choice of how to vary these answers (even-numbered visitors get

true answers) is accidental.

All this points us to clarify Calibration Reliability. We will need to look at the truth-ratio of the propositions the process actually assigned credence to, but we will also have to look at counterfactual assignments, too. What we get is the following view:

Calibration Reliability: To determine the reliability of a process ρ:

1. Construct a set of propositions that contains all the propositions that ρ has actually assigned credence to *and* all the propositions the process would assign credence to in nearby counterfactual scenarios.
2. Partition this set of propositions according to the credence assigned to them.
3. Let the score for each partition be the squared difference between the credence and the truth-ratio of propositions in that partition.
4. The weighted average of the scores for each partition is the reliability of ρ.

With the view stated, we can now work our way to responding to the worry Goldman (1986) raises for Calibration Reliability. To get to this response, consider first a specific scenario seemingly unrelated to Goldman's worry. Suppose there are a pair of processes for diagnosing whether a patient has a broken finger. Process c detects swelling and produces a credence of 0.7 in the proposition that the finger is broken. Process d detects swelling *and* the level of pain the patient is experiencing. If d detects swelling with minimal pain, it produces a credence of 0.5 in the proposition that the finger is broken. If it detects swelling with lots of pain, it produces a credence of 0.9 in the proposition that the finger is broken. Suppose that each process has been used on the same 100 patients all of whom have swelling and 70 of whom have broken fingers. Further, assume that 50 patients have minimal pain; of those, 25 have broken fingers. Finally, assume that 50 patients have lots of pain; of those, 45 have broken fingers. At first glance, d and c look like they might each come out as perfectly reliable according to Calibration

Reliability.[26] But this might seem objectionable.[27]

In answering this worry, the key thing to keep in mind is that we should not merely look at the 100 actual cases that these processes have performed on. We must also look at how they perform in nearby worlds. Which worlds are nearby depend on features of the processes, and the kinds of regularities in the environment of the actual world. In the case here, there are two important environmental scenarios to distinguish. In the first scenario, it is an accident that 70 of the 100 patients had broken fingers because it is not in general the case that 70% of those in hospitals with swollen fingers have broken fingers. What is not an accident is that 90% of those with swelling and lots of pain have broken fingers, nor is it an accident that 50% of those with swelling and minimal pain do. In the second scenario, it is similarly not an accident that 90% of those with swelling and lots of pain have broken fingers, nor is it an accident that 50% of those with swelling and minimal paint do. But in addition, it is not an accident that 70% of those in hospitals with swollen fingers have broken fingers. There are, of course, other scenarios (for instance, when all the statistics are mere accidents), but these two are all that will concern us. The important thing to notice is that if scenario 1 obtains, then c will not be perfectly calibrated. This is because c produces 0.7 credence to all the propositions about broken fingers when it detects swelling even though the truth-ratio in the set of propositions (drawn now from nearby worlds) is not 0.7. If, on the other hand, scenario 2 obtains, then both processes are perfectly calibrated and thus perfectly reliable.

This puts us in a good position to see the appropriate response to Goldman's worry. Recall, Goldman is worried that a calibration-based approach to reliability will sanction incorrect reliability evaluations, in particular when it comes to entire belief states. Suppose it is true of Lisa, that out of all the propositions she has an opinion about, 70% are true. If so, then it might seem that her credence function achieves perfect calibration by setting her credence equal to 0.7 for all propositions about which she has an opinion. But to say that such a credence function is perfectly reliable, Goldman

[26] To determine the reliability of process c, we have one partition that includes all 100 propositions. Its score is $(0.7-0.7)^2 = 0$, which means it is perfectly reliable. For Process d, we first partition the 100 propositions into two sets: a set of those propositions assigned 0.5 credence and a set of propositions assigned 0.9 credence. We then work out the score for each set (in this case, the first set's score is $(0.5-0.5)^2 = 0$ and the second set's score is $(0.9-0.9)^2 = 0$). To work out the reliability of d, we then take the weighted average of these scores, which is 0. d is thus assessed as perfectly reliable, too.

[27] I address this kind of case in more detail in section 4.2.

thinks, is clearly wrong.

The first thing to say is that assessing an entire belief state for reliability is actually a tricky matter. Consider an example from binary models of belief. If all you know is that 90% of Joe's beliefs are true, this doesn't settle the question about the overall reliability of his belief state, because you don't know enough about how he came to have those beliefs so that we can determine which counterfactual scenarios must be taken into account. Suppose Joe forms all his beliefs by asking the crazy professor, and has just happened to usually be an even-numbered visitor. In this case, although the propositions Joe actually believes have a high truth-ratio, we would be unlikely to say that his overall belief state (or Joe himself) is reliable.

With this in mind, consider Lisa. We don't just want to look at the propositions to which she has actually assigned credence. We also want to look at the propositions she assigns 0.7 credence to in nearby worlds. If it's just chance that 70% of the propositions Lisa actually assigns credence to are true, then the propositions used to evaluate her level of calibration won't have a 0.7 truth-ratio and Lisa won't come out as having a reliable belief state. Another way to think about this is that her belief state is not in a *stable* state of matching the truth-ratios. She happens to match the truth-ratio of the propositions she now believes, but if we come back and evaluate her after she has formed other credences and dropped some existing ones, there's a good chance her credences will no longer match the truth-ratio of propositions believed. Thus her belief state is not assessed as perfectly reliable.

There are some credence functions that will assign all propositions one credal value and still be in a stable state of matching the truth-ratios. Suppose that Lisa has credence in a proposition if and only if she has credence in the negation of that proposition. It is easy to see that the truth-ratio for this set of propositions will be 1/2 in *every* world. Thus, Lisa can achieve perfect calibration by setting her credence in each proposition to 0.5. It is a consequence of Calibration Reliability that this credence function is perfectly reliable. However, I do not think that this is counterintuitive. After all, a simple way of generating a reliable belief state in a binary model is to only believe propositions of the form $P \vee \neg P$. There is clearly something undesirable about such a belief state; but this does not stem from its unreliability. Suppose that I see a tiger running toward me in broad daylight. Suppose that instead of forming the belief that there is a tiger running toward me, I form the belief that either there is a tiger running toward me or there is not a tiger running toward me. There are obvious problems I'll run into if this is how I form beliefs, but the problem is not that I'm unreliable.

Note, finally, that this view can handle the problem cases that arose for Brier Reliability. Calibration Reliability rates processes a and b as perfectly reliable, and a^* as less reliable than both. Further, Calibration Reliability does not exclude processes that produce mid-level credences from having high reliability scores (as is evidenced by process c).

This concludes my presentation of Calibration Reliability. In the next section I will consider objections to the view. But before this it is worth noting that just as Brier Reliability can be seen as a generalization of reliability in binary models, so too can Calibration Reliability, albeit in a slightly different way. Suppose we think of binary belief as corresponding to credence 1. Then, when working within binary models, every process will produce belief states of the form $[c(P) = 1]$.[28] If so, there is only one cell partitioning the set of propositions believed (those assigned credence 1) and the Calibration Score for the process is solely a function of r_1, the truth-ratio of the propositions in that cell. This is the quantity to which reliabilists typically appeal: the truth-ratio of the propositions believed.[29]

4 Objections to Calibration Reliability

4.1 Calibration Reliability and Expectation

Above I criticized the use of improper scoring rules in truth-possession accounts of reliability. The objection asked us to consider an agent who seeks to maximize expected reliability. If Linear Reliability were true, then when such an agent looks at one of her belief-forming processes she will expect an alternative process to do better. This kind of second-guessing will continue until she has a process that assigns either all 0s or all 1s to propositions.

It is important to see that Calibration Reliability doesn't have this exact problem, though as we'll see, it may seem to have a related one. Focus on a process that produces only one credal value, b, to a set of propositions $\mathbf{P}$. The Calibration Score for each such credence is $(r-b)^2$ where r is the truth-ratio of the propositions. Since, there is only one credal value, b, Calibration

[28] I do not mean to commit myself to a controversial thesis about the relation between graded belief and binary belief. I simply mean that in evaluating processes for reliability, it is harmless to think of a binary belief as corresponding to a graded belief of 1. The '1' simply indicates that the proposition assigned 1 is believed.

[29] The reliability verdicts given by Calibration Reliability to processes in binary models are ordinally equivalent to those given by standard truth-ratio based verdicts. Like Brier Reliability, however, there will be some cardinal differences in the verdicts given by truth-ratio based approaches and Calibration Reliability. This is due to the fact that both the Brier and the Calibration Scores are quadratic rules, which *square* the error term.

Reliability says that this score gives us the reliability of the process. The expected Calibration Score for such a process is thus $\sum_{i=0}^{1} c(r = i) \times (i - b)^2$, where $c(r = i)$ is the credence the agent gives to the proposition that r takes value i. According to some views, the values that $c(r = i)$ can take are restricted by the credence (b) assigned to the propositions in $\mathbf{P}$. For instance, suppose that $\mathbf{P} = \{P_1, P_2\}$, $c(P_1) = c(P_2) = 0.7$, and the agent views P_1 and P_2 as independent. If r were the truth-ratio of the propositions actually believed, the agent would have to think that either $r = 1$, $r = 0$, or $r = 0.5$ with probability 0.49, 0.09, and 0.42 respectively. This need not be the case given our interpretation of r in the context of Calibration Reliability, since r is the truth-ratio of propositions actually believed on account of the process *and* those believed in nearby worlds. Suppose, then, that for some process, $c(r = k) = 1$. It follows immediately that the agent expects to be the most reliable with a process that assigns credence $b = k$ to the propositions in $\mathbf{P}$. The agent is not compelled to move to a process that assigns only extreme values to propositions and thus Calibration Reliability does not have the same unacceptable consequence as Linear Reliability.

However, there is a potential worry along these lines. Suppose that there is a process that assigns credence to a set of propositions $\mathbf{P}$ where $\mathbf{P}$ is special in that if a proposition is in $\mathbf{P}$, then so is its negation. In such a case the agent can be certain that there is an alternative process that always does at least as good as this process in terms of calibration: it is the process that assigns to every member of $\mathbf{P}$ a credence $b = 1/2$.[30] Thus, she has a weak dominance argument for preferring this process to the initial one.

How serious is this problem? For several reasons it is less serious than the problem that arises for those who appeal to something like Linear Reliability. First, note that this problem only arises for processes that issue credences in a set of propositions with the appropriate structure. So it is not as wide a problem as arises for Linear Reliability.

Second, suppose the agent is currently using a process that assigns different credal levels to the propositions in $\mathbf{P}$. For concreteness, let $\mathbf{P}_{0.8}$ be the subset of the propositions in $\mathbf{P}$ to which a credence of 0.8 is assigned and $\mathbf{P}_{0.2}$ be the remainder to which a credence of 0.2 is assigned. If the agent has what are reasonable beliefs about the truth ratios of $\mathbf{P}_{0.8}$ and $\mathbf{P}_{0.2}$—namely that $r_{0.8} = 0.8$ and that $r_{0.2} = 0.2$—then she will think her current process is doing just as well as the trivial process that would assign all the propositions

[30] This follows from the fact that $\mathbf{P}$ is constructed so that exactly half of its members are true.

in **P** a credence of 1/2. Thus, she is not *compelled* to switch to the trivial process. Again, this differs from the problem with Linear Reliability, where a reliability-maximizing agent will be compelled to change to an alternative process.

Finally, if Linear Reliability were true, a reliability-maximizing agent will not only be compelled to change to an alternative process, she will also not expect this change to have any epistemic downside. However, as we'll see in the next section, the trivial process that assigns all propositions a credence of 1/2 guarantees reliability, but comes with an epistemic cost.

4.2 The Burglar Alarm and The Broken Finger

Consider two specific cases that will illustrate the epistemic cost associated with the trivial process mentioned in the previous section. These two cases are also objections to Calibration Reliability in their own right. The first case is the comparison between processes c and d from section 3.3. Recall that c detects swelling and produces a credence of 0.7 in the proposition that the finger is broken. Process d detects swelling *and* the level of pain the patient is experiencing. If it detects swelling with minimal pain, it produces a credence of 0.5 in the proposition that the finger is broken. If it detects swelling with lots of pain, it produces a credence of 0.9 in the proposition that the finger is broken. As we saw, under some conditions Calibration Reliability does not say that the two processes are equally reliable. But if all the statistics are non-accidental, then Calibration Reliability does say that these two processes are equally reliable. The objection is that they aren't really equally reliable: imagine a doctor who is asked to choose between using c and d. The decision is clear: d will be preferred since it provides more discriminating information. This purports to show that the reliability evaluations issued by Calibration Reliability are mistaken.

Consider another case. Suppose you are designing a security system. Process e produces credences of the form $[c(\text{Burglar}) = 0.6]$ in response to a certain sound. Further, 60% of the time there is such a sound, there is a burglar. Process f, on the other hand, produces credences of the form $[c(\text{Burglar}) = 0.96]$ in response to a certain shadow. 95% of the time there is such a shadow, there is a burglar. Now, it seems that if we were designing a security system, we would prefer to have f rather than e.[31] But note that e comes out as perfectly reliable according to Calibration Reliability, whereas

[31] Such a decision depends on the probability that there is the relevant shadow given that there is a burglar, and the probability that there is the relevant sound given that there is a burglar. But if these are roughly the same, then f does seem preferable to e.

f does not. This, too, one might argue, shows that Calibration Reliability is mistaken.

My response to both cases is the same. Calibration Reliability tells us which processes are reliable, but there are epistemic virtues other than reliability. Process d is better than c and process f is better than e in some of these respects. But they are not better in terms of reliability.

An example drawn from binary models of belief can make the nature of this response clearer. Consider Process 1. It examines the patients with injured fingers and produces (full) beliefs of the form: *patient i has a broken finger or a jammed finger*. Its truth-ratio is 0.9. Process 2 produces (full) beliefs of the form: *patient i has a broken finger*. Its truth-ratio is also 0.9. It is uncontroversial that they are equally reliable. This is true despite the fact that Process 2 faces a more difficult task than Process 1 and also presents a doctor with more useful information than Process 1. Process 2 is to Process 1 as d is to c. Both Process 2 and d are more discriminating than their counterparts. Nevertheless, they are still just as reliable as their counterparts. So, I maintain that calibration is a good measure of reliability.

But perhaps we can change the example slightly to make things more troubling. Suppose that there is one doctor who has both processes c and d available to him. According to Calibration Reliability, if he forms the graded belief [c(broken) = 0.7] via process c, then this is fully justified. Similarly, if he forms the graded belief [c(broken) = 0.9] via process d, this is fully justified, too. But, the objection goes, if the doctor relies on c instead of d in a certain circumstance, the graded belief thereby formed ([c(broken) = 0.7]) would be unjustified. This goes against Calibration Reliability.[32] This kind of example takes us into difficult issues that have more to do with reliabilism about justification in general, rather than its specific application to a graded model of belief. For note that these kinds of quandaries arise for reliabilists using binary models of belief, too. Suppose that I overhear someone in the hallway of my apartment talking about it now raining in my neighborhood while I am at the same time looking online where it says it is not raining. Here we have two processes, each reliable on their own, that issue different verdicts about what to believe. Something must be said about what one is justified to believe in this situation, but doing so requires a much more elaborate theory than the simple reliabilist theory of justification with which we've been working. So I wouldn't here like to take too firm a stand on what to say about this sort of case. But it seems plausible to say in the specific case where the doctor has both c and d available to him, that he would be

[32]Thanks to an anonymous referee for giving this objection.

more justified in forming the belief in virtue of d rather than c because d uses all the evidence that c uses (the presence of swelling) and more (whether there is significant pain or minimal pain). This requires one to say something slightly more complicated about the relationship between justification and reliability, but it does not cast into doubt the identification of reliability with the Calibration Score. Again, then, I maintain that the Calibration Score is a good measure for reliability.

If that's right, then there is an interesting question about what epistemic virtue d and f have over c and e. One promising line of thought is that the Refinement Score captures the kind of epistemic virtue they have. For instance, although d and c have the same Calibration Score and so are equally reliable, d has a better Refinement Score.[33] Similarly, although e is just barely more calibrated and thus more reliable than f, f has a much better Refinement Score. The Brier Score then allows us to combine the virtues of calibration (reliability) and refinement.

What is the epistemic virtue that the Refinement Score is capturing? A process has a good Refinement Score when most of the propositions it assigns one credal value to are either all true or all false. That is, a process that can group propositions into sets that are mostly true or mostly false—and then assign those propositions different credences—is one that will get a good Refinement Score. Thus, the Refinement Score seems to correspond to something like discriminatory ability or potential informativeness of a process. If a process has this ability *and* is also highly reliable, then the process will assign very high credences to propositions that are mostly true and very low credences to propositions that are mostly false.

One interesting thought is that the Refinement Score might capture something like what Goldman (1986) has called the *power* of a process. According to Goldman: "Power is the capacity of a process, method, system, or what have you to produce a large number of true beliefs; or, slightly differently, the capacity to produce true beliefs in answer to a high ratio of questions one wants to answer or problems one wants to solve." (Goldman, 1986, p. 26) Later in the book, Goldman notes that cognitive systems low in power leave the cognizer with very little information about questions the cognizer wants answered. (p. 122) Here, two properties are being equated: the property of producing very few beliefs in answer to questions of interest and the property of leaving the cognizer with very little information with respect to questions of interest. Goldman suggests that power is a measure

[33] If we stick to the 100 propositions actually assigned credence, $R(d) = 0.17$ while $R(c) = 0.21$.

of these properties. In graded models, however, these two properties can come apart. One way for a process to generate very little information with respect to a set of questions is for it to produce very few graded beliefs in response to those questions. Another way for a process to generate very little information with respect to a set of questions is for it to produce graded beliefs that assign uninformative credences to propositions that are answers to those questions. The Refinement Score of a process, if it measures anything like the power of a process, will measure power in this latter sense.

Though there is much more to say about this, a simple example displays the basic point. Consider a reliable process—that is, a highly calibrated process—but with very little power as measured by the Refinement Score. Since the process is highly calibrated, the truth-ratio, r_j of propositions assigned credence b_j ($0 \leq b_j \leq 1$) is such that $r_j \approx b_j$. If the Refinement Score for this process is nevertheless poor, this means that for most j, $(r_j)(1 - r_j)$ is high, which happens when r_j is near $1/2$. Accordingly, a low-power, high-reliability process is one that assigns credences near $1/2$ to most propositions. Since in many contexts such credences are uninformative, such a low-power process will generate relatively uninformative credences.[34]

As indicated above, there is much more to say about how exactly to understand the Refinement Score and how it might relate to Goldman's informal notion of power. But in identifying reliability with the calibration of a process, it opens up the interesting possibility of investigating the other component of the Brier Score and looking at its epistemic importance. It also shows that valuing reliability may not be in tension with valuing truth-possession. Rather, reliability might itself be a means to truth-possession.

[34]One might note that sometimes mid-level credences are informative. For example, suppose there are 11 possible answers to a question among which the inquirer is indifferent. After applying the process in question one answer is assigned credence 0.5 and the rest are assigned credence 0.05. This seems to be informative. Does this show that the Refinement Score doesn't really measure informativeness? No. To see this note that there will be two cells in the partition for this process. In the first, $\mathbf{P}_{0.05}$, are all the propositions assigned credence 0.05. This set is 10 times larger than the set of propositions, $\mathbf{P}_{0.5}$, which are assigned credence 0.5. If we assume that the process is perfectly reliable (that is, perfectly calibrated), then the the truth-ratios for these sets of propositions are $r_{0.05} = 0.05$ and $r_{0.5} = 0.5$. Thus, the Refinement Score is approximately 0.066 for this process, which is fairly good. Why is this? Although the credence of 0.5 is not all that informative, most propositions are assigned the informative credence of 0.05. Thus, in a simple case where we think a mid-level credence *is* informative, the Refinement Score tracks this.

4.3 Why Care?

I will consider one final objection: if the sort of reliability that is defined by Calibration Reliability doesn't track truth-possession and it doesn't track the kind of discriminating power or informativeness that doctors might care about when diagnosing patients, then why care about it at all? I have claimed that the kind of reliability that Calibration Reliability gets at is the sort of reliability closely aligned with justification for reliabilists. But perhaps the focus on justification is itself a mistake. Perhaps this kind of reliability is just uninteresting.

There are several things that can be said in response to this sort of concern. The first might be obvious from what was said immediately above: even if reliability isn't the *only* thing we care about when evaluating a belief-producing process, it is certainly one component that we care about. In addition, by isolating the reliability of a process from its other epistemic virtues, we get a clearer picture of what those other epistemic virtues might be.

Second, we often care about the justificatory status of beliefs. For example, Hume's problem of induction is unsettling precisely because the conclusion states that none of our beliefs about the future are justified. If we think that belief comes in degrees, then Hume's conclusion is that our credences about the future are not justified. One might not like reliabilist responses to Hume's problem, but without an understanding of the kind of reliability that is closely related to justification, we cannot even understand such responses when working with graded models. Consider a related example. Suppose a juror decides that she will only believe the witness's testimony if the witness is justified in believing what he testifies. Suppose that the witness reports his credence that the assailant had brown hair. What is it for this credence to be justified? If one is a reliabilist, then to answer this question one needs to know what it is for a credence to be reliably produced. Notice, further, that if Brier Reliability were correct, then less extreme levels of confidence will tend to be less justified. But that seems wrong. A report of 60% confidence that the assailants hair was brown could be highly justified. This is a further reason for interest in the kind of reliability that is defined by Calibration Reliability.

Perhaps an even more compelling response to the concern, however, is to illustrate a specific area within graded models of belief where we need to appeal to reliability, and where this reliability must be independent of truth-possession and informativeness. I'll give one example where this is the case.

One important part of Bayesian epistemology is to explain how credences are updated over time upon the receipt of new evidence. The traditional view about this is that updating proceeds via *Conditionalization*.[35] One important feature of this updating rule is that an agent's evidence propositions are assigned credence 1. An important issue for this kind of model, then, is to say which propositions represent an agent's evidence and why. Several authors have addressed this issue.[36] As is well known, however, Richard Jeffrey (1965) presents a modified framework where evidence propositions need not receive credence 1. According to this view, evidence—like belief—comes in degrees. It is equally important on this kind of model to say what an agent's degreed evidence is and why.[37]

One interesting thought is that information about *reliable* processes could be useful in addressing this issue. We could say that those processes that are highly reliable are the processes that yield evidence. For instance, one might say that if a highly reliable process produces a credence of 0.7 in S, then one's evidence at that time is $\{\langle S, 0.7\rangle, \langle \neg S, 0.3\rangle\}$. If we go this way, then we need a way of understanding reliability that is independent of either truth-possession or informativeness. To see this, suppose there is a process of shape-recognition that is highly calibrated, but does not deliver credences very close to 1. A typical output might be something like [c(object is square) = 0.7]. Such a process could be highly reliable, and thus deliver evidence, even though the credences it produces are not very close to 1. In such a case we need an understanding of reliability that allows that reliability can come apart from truth-possession. If we look only at degree of truth-possession we will build in a bias towards processes that produce high credence values in certain propositions. In this context, however, we want to allow evidence that is graded, but not necessarily close to 1. Similarly, we would like to allow that an agent has evidence-yielding processes, even if they are not highly informative. Of course, an agent that has more informative ways of gathering evidence may end up doing *better* than an agent that has less-informative ways of gathering evidence. But this doesn't

[35] Where $c(\cdot)$ is an agent's current credence function, and $c^E(\cdot)$ is that agent's credence function after learning evidence E (and nothing else), Conditionalization says that the following should hold:

COND: For all A, E, $c^E(A) = c(A|E)$ (so long as $c(E) \neq 0$).

[36] See, for instance, Maher (1996), Williamson (2000), Bird (2004), Silins (2005), Neta (2008), Dunn (2012).

[37] Jonathan Weisberg (2009) has dubbed this problem the *Inputs Problem*. Williamson (2000) rests his rejection of Jeffrey's framework on the difficulty he sees in solving—or even saying anything useful about—the Inputs Problem. See Weisberg (2011) for further discussion of why the Inputs Problem is important.

change the fact that both agents may have highly reliable processes that do yield them evidence.

5 Conclusion

In binary models of belief, reliability is uncontroversially understood in terms of truth-ratios. When we move to graded models, however, there are two different ways to understand reliability. One approach focuses on degree of truth-possession, the other focuses on degree of calibration. I've argued that focusing on degree of calibration, and not degree of truth-possession, is the best way to generalize the concept of reliability to these graded models. If that's right then it shows several things of interest. First, it shows that the most natural proposal for how to generalize reliabilism to graded models is mistaken. Second, it shows that reliability is not *essentially* connected with high levels of truth-possession; rather, this is an accidental feature of reliability that arises only when working with binary models of belief. This has important ramifications for discussions about epistemic value. Third, it points towards an interesting project of cataloguing and better understanding the epistemically worthwhile properties that processes of belief production can have. Goldman himself initiated this project in *Epistemology and Cognition*, distinguishing between reliability, power, and speed. The arguments here suggest that in looking at graded models of belief we might be able to make progress in drawing the distinctions necessary for an overall epistemic evaluation of a process. Fourth, and finally, I have suggested that reliability, as defined by Calibration Reliability, may be important in saying something about evidence within certain Bayesian models.[38]

References

Armstrong, D. (1973). *Belief, Truth and Knowledge*. Cambridge: Cambridge University Press.

Baumann, P. (2009). Reliabilism—modal, probabilistic, or contextualist. *Grazer Philosophische Studien*, *79*, 77–89.

[38]Earlier versions of this paper were given at the Fall 2011 Meeting of the Indiana Philosophical Association, the 2012 Central APA, and at Western Michigan University. Thanks to all participants there. Thanks especially to Erik Wielenberg, Jennifer Lackey, Lara Buchak, Chris Meacham, James Joyce, Ethan Brauer, and anonymous reviewers for *Philosophical Studies* for very helpful comments.

Bird, A. (2004). Is evidence non-inferential? *The Philosophical Quarterly*, *54*, 252–265.

Blattenberger, G. & Lad, F. (1985). Separating the Brier score into calibration and refinement components: A graphical exposition. *American Statistician*, 26–32.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Cohen, S. (1984). Justification and truth. *Philosophical Studies*, *46*, 279–295.

De Finetti, B. (1972). *Probability, Induction and Statistics: The Art of Guessing*. New York: Wiley.

DeGroot, M. & Fienberg, S. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *32*(1/2), 12–22.

DePaul, M. (2004). Truth consequentialism: Withholding and proportioning belief to the evidence. *Philosophical Issues*, *14*(1), 91–112.

Dunn, J. (2012). Evidential externalism. *Philosophical Studies*, *158*(3), 435–455.

Gibbard, A. (2008). Rational credence and the value of truth. In T. Gendler & J. Hawthorne (Eds.), *Oxford Studies in Epistemology*, Oxford: Oxford University Press, vol. 2.

Goldman, A. (1979). What is justified belief? In G. Pappas (Ed.), *Justification and Knowledge*, Dordrecht: D. Reidel. 1–23.

Goldman, A. (1986). *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goldman, A. (1988). Strong and weak justification. *Philosophical Perspectives*, *2*, 51–69.

Goldman, A. (1999). *Knowledge in a Social World*. Oxford: Oxford University Press.

Goldman, A. (2010). Epistemic relativism and reasonable disagreement. In R. Feldman & T. Warfield (Eds.), *Disagreement*, Oxford: Oxford University Press. 187–215.

Goldman, A. (2012). *Reliabilism and Contemporary Epistemology*. Oxford: Oxford University Press.

Goldman, A. & Shaked, M. (1991). An economic model of scientific activity and truth acquisition. *Philosophical Studies*, *63*(1), 31–55.

Greaves, H. & Wallace, D. (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind*, *115*(459), 607–632.

Greco, J. (1999). Agent reliabilism. *Noûs*, *33*, 273–296.

Hájek, A. (unpublished). A puzzle about degree of belief. `http://fitelson.org/coherence/hajek_puzzle.pdf`.

Jeffrey, R. (1965). *The Logic of Decision*. Chicago: University of Chicago Press.

Joyce, J. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *65*(4), 575–603.

Joyce, J. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of Belief*, Springer. 263–297.

Lange, M. (1999). Calibration and the epistemological role of Bayesian conditionalization. *The Philosophical Review*, *96*(6), 292–324.

Leitgeb, H. & Pettigrew, R. (2010a). An objective justification of Bayesianism I: Measuring inaccuracy. *Philosophy of Science*, *77*(2), 201–235.

Leitgeb, H. & Pettigrew, R. (2010b). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science*, *77*(2), 236–272.

Maher, P. (1996). Subjective and objective confirmation. *Philosophy of Science*, *63*, 149–174.

Murphy, A. (1972). Scalar and vector partitions of the probability score: Part I. two-state situation. *Journal of Applied Meteorology*, *11*, 273–282.

Murphy, A. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*, 595–600.

Neta, R. (2008). What evidence do you have? *British Journal for the Philosophy of Science*, *59*, 89–119.

Pritchard, D., Millar, A., & Haddock, A. (2010). *The Nature and Value of Knowledge: Three Investigations*. Oxford: Oxford University Press.

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The Foundations of Mathematics and other Logical Essays*, Routledge. 156–198.

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, *2*(2), 191–201.

Savage, L. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.

Schmitt, F. F. (2000). Veritistic value. *Social Epistemology*, *14*(4), 259–280.

Seidenfeld, T. (1985). Calibration, coherence, and scoring rules. *Philosophy of Science*, *52*, 274–294.

Silins, N. (2005). Deception and evidence. *Philosophical Perspectives*, *19*, 375–404.

van Fraassen, B. (1983). Calibration: A frequency justification for personal probability. *Boston Studies in the Philosophy of Science*, *76*, 295–319.

Weisberg, J. (2009). Commutativity or holism? A dilemma for conditionalizers. *The British Journal for the Philosophy of Science*, *60*(4), 793.

Weisberg, J. (2011). Varieties of Bayesianism. In D. Gabbay, S. Hartmann, & J. Woods (Eds.), *Handbook of the History of Logic*, Elsevier, vol. 10.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford: Oxford University Press.