Check for updates

DOI: 10.1111/phpe.12156

ORIGINAL ARTICLE

PHILOSOPHICAL PERSPECTIVES WILEY

Deference Done Better

Kevin Dorst[1,2,*] | **Benjamin A. Levinstein**[3,*] | **Bernhard Salow**[1,*] | **Brooke E. Husic**[4,†] | **Branden Fitelson**[5,†]

[1] University of Oxford
[2] University of Pittsburgh
[3] University of Illinois at Urbana-Champaign
[4] Stanford University
[5] Northeastern University

Correspondence
Kevin Dorst, University of Pittsburgh.
Email: kevindorst0@gmail.com

[*] K.D., B.A.L., and B.S. contributed equally—they jointly proved the theorems and wrote the paper.
[†] B.E.H. and B.F. pioneered the computational methods and located the key counterexample (Fact 2.1); together, these made the rest of the paper possible.

Abstract

There are many things—call them 'experts'—that you should defer to in forming your opinions. The trouble is, many experts are *modest*: they're less than certain that they are worthy of deference. When this happens, the standard theories of deference break down: the most popular ("Reflection"-style) principles collapse to inconsistency, while their most popular ("New-Reflection"-style) variants allow you to defer to someone while regarding them as an *anti*-expert. We propose a middle way: deferring to someone involves preferring to make any decision using their opinions instead of your own. In a slogan, *deferring opinions is deferring decisions*. Generalizing the proposal of Dorst (2020a), we first formulate a new principle that shows exactly how your opinions must relate to an expert's for this to be so. We then build off the results of Levinstein (2019) and Campbell-Moore (2020) to show that this principle is also equivalent to the constraint that you must always expect the expert's estimates to be more accurate than your own. Finally, we characterize the conditions an expert's opinions must meet to be worthy of deference in this sense, showing how they sit naturally between the too-strong constraints of Reflection and the too-weak constraints of New Reflection.

 © 2021 Wiley Periodicals LLC

Some people are fools. It's best to ignore them in forming your opinions—and from here on out, we will.

Others are not. In fact, there are many things—call them 'experts'—that it's best to *defer to* in forming your opinions. This paper offers a theory of how.

Experts include people who are smarter and better informed than you, but also things like your current evidence and the objective chances. Although such experts are clearly worthy of deference in some sense, it's tricky to say exactly what that sense is. The trouble is that it's possible for experts to be *modest*: to be less than certain that their opinions are indeed worthy of deference.[1] The person who is in fact smarter and better informed than you might not be sure about that. Your evidence might in fact support q, but also assign some probability to the claim that your evidence doesn't support q. And the objective chances (at least on certain Humean theories) might assign non-zero chance to *undermining futures*: futures such that, if they came about, the objective chances would now have been different from what they actually are. When any of these things happen, the standard theory of deference—which is built around "Reflection" principles—breaks down.[2] A better theory is needed.[3]

This paper offers one. Building on Dorst (2020a), we suggest that the right deference principle should encode the idea—call it 'Value'—that you defer to an expert when you'd prefer to make any decision using their opinions instead of your own (cf. Huttegger 2014; Schoenfield 2016b; Nissan-Rozen and Spectre 2019). As we'll sometimes say: deference is a preference to give someone power of attorney. In a slogan: *deferring opinions is deferring decisions*.

Unlike Reflection, Value allows modesty. Nevertheless, it turns out to be a surprisingly strong constraint. The most popular way of fixing Reflection to allow modesty ("New-Reflection"[4]) allows for radical failures of Value. And we'll see that even deference understood along the precise lines suggested by Dorst (2020a) himself ("Trust") does not entail Value in full generality. But his suggestion can be modified to yield a deference principle ("Total Trust") which *is* exactly equivalent to Value—and which, we argue, offers an account of deference that's better than those currently on offer.

Moreover, although this account was built for modest experts, it turns out to have implications for deference to *im*modest ones as well. This may be surprising, as it's well known that when experts are immodest, Value is equivalent to Reflection (Skyrms 1990; Huttegger 2014). However, that result holds only because the standard accounts of deference—including our statement of Value above—are accounts of *global* deference: of how to defer to an expert on *every* question. This type of deference is appropriate for some experts (the objective chances, your current evidence), but not for others: you defer to the weather forecaster *about tomorrow's weather*, but not about whether you should throw out the bananas you bought last weekend. Thus we also need a notion of *local* deference: deferring to an expert's opinions on certain questions, but not others.

To formalize local deference, we can relativize our principles to a given question Q—for example, you Value an expert *with respect to tomorrow's weather* iff, for any decision *whose outcomes are determined by tomorrow's weather*, you'd like to give them power of attorney on that decision. In a surprising twist, it turns out that Value and Reflection are *not* equivalent when relativized to questions in this way, even when the expert is immodest (§5). Thus our argument that Value is the correct account of (global) deference to modest experts suggests that it may also have an important role to play in understanding (local) deference more generally. Nevertheless, our hands will be plenty full just getting a theory of global, modest deference on the table—so we'll largely suppress talk of local deference until the end.

Here's the plan. §1 sharpens both the problem raised by modest experts and our proposal for how to fix it. §2 generalizes Dorst's deference principle and proves that this generalization is

indeed equivalent to Value. §3 then explores how Value—which, with its focus on decisions, can seem like a *pragmatic* deference principle—relates to a purely epistemic variant of the same idea. Building on the results of Levinstein (2019) and Campbell-Moore (2020), we show that Value is equivalent to *Epistemic* Value: the claim that you should expect the expert's estimates to be more accurate than your own on any reasonable way of measuring their accuracy. §4 then characterizes the conditions the potential experts' opinions must meet in order to be worthy of deference in our sense, comparing them to those imposed by Reflection and New Reflection. §5 concludes by turning to local deference and several other open questions.

One theme of this paper is that allowing experts to be modest opens up a largely unexplored domain of potential deference principles. In an attempt to help make this domain more tractable, we've included a Mathematica notebook that contains functions for generating, visualizing, and exploring the models of modesty used throughout.[5]

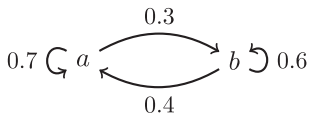
Finally: the goal of this paper is to develop and defend a philosophical idea about deference. That defense is built upon a series of equivalence and non-equivalence results, which in turn require a fair bit of technical exposition to establish. Nevertheless, we hope that those uninterested in the technicalities will care about the philosophy. So for those who care about deference but not about hyperplanes, we've done our best to hide the technicalities in endnotes, "*Proof*" paragraphs, and appendices—these can (and perhaps should) be skipped without loss to the philosophical argument.

1 | DEFERRING OPINIONS AS DEFERRING DECISIONS

Fix on some particular expert—say, your current evidence. Let 'P' be a definite description for 'the expert's probabilities, whatever they are'. There are various possibilities for what those probabilities might be, which we model with a **probability frame** $\langle \boldsymbol{W}, \boldsymbol{\mathcal{P}} \rangle$.[6] (For introductions to how models of this kind work, see e.g. Williamson 2008, 2019; Dorst 2019b, 2020b.) '$\mathcal{P}_w$' is a rigid designator for the probability distribution the expert has at world w.[7] Propositions are modelled as subsets of W. Using a standard move from modal logic (Hintikka 1962; Kripke 1963), this includes propositions about the expert's probabilities: the proposition that the expert assigns probability t to q, written '$[P(q) = t]$', is simply the set of worlds w such that $\mathcal{P}_w(q) = t$; the proposition that the expert has probability function $\mathcal{P}_x$, written '$[P = \mathcal{P}_x]$', is simply the set of w such that $\mathcal{P}_w = \mathcal{P}_x$; and so on. For the expert to be modest at w is for them to assign non-zero probability to 'the expert's probabilities' (picked out non-rigidly) being different from their own (picked out rigidly).[8]

Here's a simple example. There are two possibilities, $W = \{a, b\}$. P represents rational credences. The evidence you have is ambiguous, so it's rational to be unsure what the rational response to the evidence is (Christensen 2010; Elga 2013; Lasonen-Aarnio 2013, 2015; Carr 2019b; Dorst 2019b, 2020a). At world a it's rational to be 70% confident you're at world a; at world b, it's rational to be only 40% confident of this. Thus $\mathcal{P}_a(a) = 0.7$ while $\mathcal{P}_b(a) = 0.4$. There are two compact ways to represent such a frame. The first is a *Markov diagram* in which nodes represent worlds, and the arrow from node w to node v is labelled with the probability $\mathcal{P}_w$ assigns to v, $\mathcal{P}_w(v)$. The second represents the frame as a (row-)*stochastic matrix*, in which row i, column j represents $P_{w_i}(w_j)$. These two representations are given in Figure 1.

This frame represents a modest expert since, for example, at a the rational credence function is unsure whether it's rational to assign credence 0.7 or 0.4 to a: $\mathcal{P}_a(a) > 0$ and $\mathcal{P}_a(b) > 0$, thus $\mathcal{P}_a(P(a) = 0.7) > 0$ and $\mathcal{P}_a(P(a) = 0.4) > 0$.



$$\begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

FIGURE 1 A modest frame, in Markov-diagram (left) and stochastic-matrix (right) forms

A probability frame represents various hypotheses about what the expert's probabilities might be. To formulate deference principles, we need to add a further probability distribution $\boldsymbol{\pi}$ over W, which represent the opinions of the individual deferring to the expert.[9] While 'P' is a definite description for 'the expert probability function, whatever it is', 'π' (along with other lowercase Greek letters, like 'ρ') is a rigid designator for a particular probability function whose values are fixed and known.

Structure in place, let's consider some deference principles: principles attempting to specify how π and $\langle W, \mathcal{P}\rangle$ need to be related for π to be deferring to $\langle W, \mathcal{P}\rangle$. The most natural such principle is that to defer to the expert is to adopt their opinions as your own, conditional on any given hypothesis about what those opinions might be:[10]

Reflection: $\pi(\cdot|P = \rho) = \rho$
Conditional on the expert having a certain set of opinions, adopt those opinions.

Let us say that π **reflects** a frame $\langle W, \mathcal{P}\rangle$ iff it meets this condition.

Unfortunately, the conditions imposed by Reflection are too strong. In particular, it's impossible for any distribution to reflect a frame while assigning non-zero probability to the possibility that the expert is modest (Hall 1994; Lewis 1994; Elga 2013). The reason is simple enough. Suppose you leave open that a modest candidate $\mathcal{P}_w$ might be the expert. Since it's modest, it assigns less than maximal credence to the claim that it's the expert: $\mathcal{P}_w(P = \mathcal{P}_w) < 1$. But conditional on it being the expert, you should be certain that it's the expert: $\pi(P = \mathcal{P}_w|P = \mathcal{P}_w) = 1$. Therefore conditional on this fact, you don't adopt the opinions of $\mathcal{P}_w$: $\pi(\cdot|P = \mathcal{P}_w) \neq \mathcal{P}_w$. For example, in the frame from Figure 1, $\mathcal{P}_a(P = \mathcal{P}_a) = 0.7$, but $\pi(P = \mathcal{P}_a|P = \mathcal{P}_a) = 1$.[11]

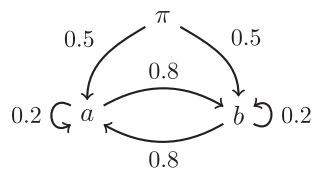
Thus if Reflection were the correct theory of deference, it would be impossible to defer to an expert unless you're sure they're immodest. But deferring to someone while leaving open that they might be modest is clearly possible. So reflecting a frame isn't necessary for π to count as deferring to it—Reflection cannot be the full theory of deference.

In response to this problem, several philosophers have argued that we should modify Reflection as follows:

New Reflection: $\pi(\cdot|P = \rho) = \rho(\cdot|P = \rho)$
Conditional on the expert having a certain set of opinions, adopt the opinions they would have were they to learn that fact.

Say that π **new-reflects** a frame iff this condition holds. New Reflection is equivalent to Reflection when the expert is certain to be immodest (since then for all w, $\mathcal{P}_w = \mathcal{P}_w(\cdot \mid P = \mathcal{P}_w)$); but New Reflection permits deference to modest experts.

Moreover, defenders of New Reflection have an attractive story to tell about why this principle should hold even when Reflection fails (Hall 1994; Lewis 1994; Elga 2013). The idea is that the condition that $P = \mathcal{P}_w$ gives additional information—information that the expert, being modest, doesn't have. So when you reason hypothetically on that assumption, you shouldn't adopt the expert's *un*conditional opinions, which fail to take this information into account; rather,



$$\pi = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$$

$$\langle W, \mathcal{P} \rangle = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}$$

FIGURE 2 A frame unworthy of deference. Left: Markov diagram. Right: stochastic-matrix

you should adopt the opinion the expert would have once made aware of this information, i.e. $\mathcal{P}_w(\cdot \mid P = \mathcal{P}_w)$.

This is a compelling explanation for why Reflection is not required for deference to experts who might be modest; and it also makes clear why, by contrast, New Reflection really is necessary for such deference. But it's worth noting that it provides no direct reason to think that obeying New Reflection is *sufficient* for deference. Moreover, there's good reason to think that it's not. The problem is that—even when the expert knows everything that you do—New Reflection imposes virtually no constraints on what you think about the expert's unconditional opinions.

To see this, notice that we can equivalently state New Reflection as follows (Stalnaker 2019). Let the **informed expert** opinions, $\widehat{\boldsymbol{P}}$, be the opinions the expert would have were their modesty to be removed—i.e. were they to be informed that they are the expert: $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{w}} := \mathcal{P}_w(\cdot|P = \mathcal{P}_w)$. Then New Reflection is equivalent to:

New Reflection (informed version): $\pi(\cdot|\widehat{P} = \rho) = \rho$
Conditional on the *informed* expert having a certain set of opinions, adopt those opinions.[12]

New Reflection thus says simply to reflect the *informed* expert, and is silent on what you should think about the expert's unconditional opinions.

This is a problem. It means that you can new-reflect a frame even if you are certain that the expert is mistaken. Consider a toy example from Dorst (2020a, §6), represented in Figure 2 below. W contains only two possibilities, a and b, and $\mathcal{P}_a(b) = \mathcal{P}_b(a) = 0.8$: both candidates for the expert assign higher probability to the false hypothesis about which of the two is the expert than the non-expert does. (For example, if a is actual, then the expert, $\mathcal{P}_a$, assigns this a low probability of 0.2, while the non-expert, $\mathcal{P}_b$, assigns it a high probability of 0.8.) Thus, for instance, conditional on the expert's credence in a being high, a is definitely false; and conditional on that credence being low, a is definitely true—the expert's credences are (known to be) anti-correlated with the truth. Meanwhile, suppose π is 50-50 between a and b, so it doesn't know anything that $\mathcal{P}_a$ and $\mathcal{P}_b$ don't. (All three are sure of $\{a, b\}$, and nothing stronger.) Surely, if π captures your credences, you do not count as deferring to this frame—the opinions (as you see them) are those of an *anti*-expert, not an expert. But note that $[P = \mathcal{P}_a] = \{a\}$ and $[P = \mathcal{P}_b] = \{b\}$, so π new-reflects this frame.[13] Upshot: π new-reflecting a frame cannot be sufficient for π to count as deferring to it; and so New Reflection cannot be the full theory of deference.

(What are we doing when we give a purely formal example like this? After all, on some interpretations of π and P, there may be general explanations for why structures like Figure 2 are impossible—for instance, perhaps the correct Humean account of objective chance will rule out the possibility of the chances having this structure (Levinstein 2019). Nevertheless, on other interpretations of 'P' (such as 'the fool's credences'), this scenario is perfectly possible. And more generally, the point of this example is to illustrate that New Reflection is missing something: the

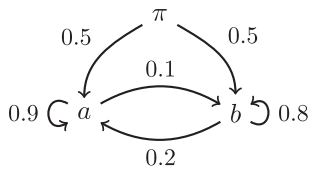
complete theory of deference should predict that π doesn't defer to P if it has this structure, yet New Reflection fails to do so.)

Intuitively, what goes wrong in the above example is that π does not regard the expert's (unconditional) opinions as tracking the facts. There is both a pragmatic and a purely epistemic version of this complaint. The pragmatic version points out that π does not regard the expert's opinions as good ones to use in making decisions: faced with a bet on which of a or b is actual, π would rather decide for itself than let the expert decide. After all, π knows that the the expert will make the wrong decision—that they'll bet on a iff $P(a) > P(b)$, which happens iff $P = \mathcal{P}_b$, i.e. iff b is actual and it's a losing bet. Conversely, they'll bet on b iff a is actual. Either way, they'll take the wrong bet—and whatever π thinks of itself, it can't think that it's bound to take the wrong bet.[14] Meanwhile, the epistemic version of this complaint points out that π regards itself as closer to the truth about which of a or b is actual than the expert: if a is actual, then π is more confident of it than the expert is ($\pi(a) = 0.5 > 0.2 = \mathcal{P}_a(a)$), and if b is actual then π is more confident of *it* than the expert is ($\pi(b) = 0.5 > 0.2 = \mathcal{P}_b(b)$).

Let's focus (till §3) on the pragmatic version. This suggests a new constraint on what deference requires: deference in opinions requires deference in decisions. That is, deference to an expert requires that, for any decision problem, you would prefer to give the expert power of attorney: to use their probabilities to make the decision rather than use your own. As we'll put it: you *value* an expert's opinions when you always expect the decisions those opinions warrant to be better than your own.

Here's how to formalize this idea. (If it's intuitive enough, skip the next two paragraphs.) Let an **option** be a function from W to real numbers, with $\boldsymbol{O(w)}$ representing the utility that would be achieved (for whatever person we are modeling with π) if O were to be chosen at w. Let a **decision problem** on W be a finite set of options $\mathcal{O}$. For example, to represent the decision of whether to bet on a or b, let $\mathcal{O} = \{O_a, O_b\}$ where $O_a = \begin{cases} 1 & \text{if } a \\ -1 & \text{if } b \end{cases}$, and $O_b = \begin{cases} -1 & \text{if } a \\ 1 & \text{if } b \end{cases}$. Relative to a probability distribution ρ, we can calculate the *expected* value of an option O as a weighted average of the various values it might take, with weights determined by how likely it is to take them: $\mathbb{E}_\rho(\boldsymbol{O}) := \sum_w \rho(w)O(w)$ (we write $\boldsymbol{E_w}(\boldsymbol{O})$ to abbreviate $\mathbb{E}_{\mathcal{P}_w}(O)$). For example, in our above anti-expert frame (Figure 2), $\mathbb{E}_\pi(O_a) = 0.5(1) + 0.5(-1) = 0$, while $\mathbb{E}_a(O_a) = 0.2(1) + 0.8(-1) = -0.6$ and $\mathbb{E}_a(O_b) = 0.2(-1) + 0.8(1) = 0.6$.

Given a frame $\langle W, \mathcal{P}\rangle$ and a decision problem $\mathcal{O}$ on W, let a **strategy** $\boldsymbol{S}$ be a way of choosing options based on the expert's probabilities: a function from W to $\mathcal{O}$, $w \mapsto S_w$ such that $S_w = S_v$ whenever $\mathcal{P}_w = \mathcal{P}_v$. A strategy S is **recommended** by the (expert modeled by the) frame for the decision-problem $\mathcal{O}$ iff the option S_w it selects at each world w is one that maximizes expected utility according to the the expert's credences at w: $\mathbb{E}_w(S_w) \geq \mathbb{E}_w(O)$ for every $O \in \mathcal{O}$. In other words, if a strategy is recommended, then following it is simply letting the expert decide how to respond to $\mathcal{O}$ on your behalf. For example, note that $\mathbb{E}_a(O_b) = 0.6 > -0.6 = \mathbb{E}_a(O_b)$, and similarly $\mathbb{E}_b(O_a) = 0.6 > -0.6 = \mathbb{E}_b(O_b)$. Thus the strategy that the frame in Figure 2 recommends for $\mathcal{O}$ is to take the bet on b (O_b) at world a, and the bet on a (O_a) at world b. As we can see, a strategy involves taking different options at different worlds. Abusing notation ever-so slightly, we'll write $\mathbb{E}_\pi(\boldsymbol{S})$ for the expected utility of following the strategy according to π: $\mathbb{E}_\pi(S) := \sum_w \pi(w)S_w(w)$. Note that in our example, this value is negative, since π knows that the expert, whoever it is, recommends taking the wrong bet: $\mathbb{E}_\pi(S) = \pi(a)S_a(a) + \pi(b)S_b(b) = 0.5O_b(a) + 0.5O_a(b) = -1$. This is less than the expected value of simply betting on a, come what may: $\mathbb{E}_\pi(O_a) = 0$. Thus π



$$\pi = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$$

$$\langle W, \mathcal{P} \rangle = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}$$

FIGURE 3 A probability function that values a modest frame

prefers *not* to give the (anti-)expert power of attorney in the above frame—it's better off choosing for itself!

Generalizing, the proposal is that π defers to a frame iff it **values** that frame:

Value: If S is recommended for $\mathcal{O}$, then for all $O \in \mathcal{O}$: $\mathbb{E}_\pi(S) \geq \mathbb{E}_\pi(O)$.
For any decision problem, prefer giving the expert power of attorney to decide on your behalf, rather than deciding for yourself.

In other words: no matter what decision you face, the expected utility of adopting a strategy recommended by the expert is always at least as high as bypassing the expert and picking an option using your own probabilities.[15]

That's the proposal. How does it relate to Reflection and New Reflection?

If π reflects a frame, it values it.[16] And if a frame is immodest, π reflects it *if and only if* π values it.[17] But this is not true in general: π can value a frame without reflecting it—and, in particular, without that frame being immodest. To get a feel for why, consider a case in which π is 50-50 between a and b but $\mathcal{P}_a(a) = 0.9$ and $\mathcal{P}_b(b) = 0.8$, in Figure 3. This frame is modest, so π doesn't reflect it—for example, $\pi(a|P = \mathcal{P}_a) = 1 \neq 0.9 = \mathcal{P}_a(a)$. But π values the frame. For if a is true then the expert is more confident of it than π is; and if b is true, likewise. As a result, P is closer to the truth than π for every proposition, and at every world. Thus, obviously, π would prefer to give the expert power of attorney. Since this expert is modest, Value must be a weaker constraint than Reflection.[18]

Meanwhile, Value is a *stronger* constraint than New Reflection. π values a frame only if it new-reflects it.[19] But, as we've seen, π does not value frames in which the expert's opinions are anti-correlated with the truth, even when it new-reflects them (Figure 2).

Finally, note that Value is equivalent to a natural formulation of a ban on Dutch-bookability. In particular, imagine that π is your prior and P is the posterior credences you will have after some transition in beliefs. A **fixed-option Dutch book** is a pair of decision problems—both including a "no bet" option with 0 payout, one presented before and the other presented after the belief-transition—such that doing the rational thing before and after is guaranteed to result in a loss.[20] If π fails to value P, it is possible to construct a fixed-option Dutch book against the transition from π to P.[21] No such book can be made if π values P; in fact, there can't even be a pair of decision problems which both include a "no bet" option but result in an *expected* loss.[22]

Value thus looks like a plausible candidate for threading the needle between the overly strong Reflection and the overly weak New Reflection. What we need now is a clearer picture of what Value requires. In particular, we'd like a characterization of Value that tells us directly how π and P's opinions must relate (§2), that explains how the pragmatic-looking Value relates to purely epistemic forms of deference (§3), and that tells us what the expert's opinions must be like in order to be valuable (§4). Off to it.

2 | IMPROVING TRUST, GENERATING VALUE

As a deference principle, Value is cumbersome—we'd like to be able to say, directly, how π's opinions must relate to P's in order for π to value P. Dorst (2020a) takes a step in the right direction. At a first-pass, he proposes as a deference principle:

Simple Trust: $\pi(q \mid P(q) \geq t) \geq t$
Conditional on the expert being confident of q, be confident of q.

This principle looks independently appealing. It says that you regard the information that the expert favors q to at least a certain extent as also favoring q to at least that extent. Note that this is not affected by the criticism of Reflection we discussed earlier. For, while it may be that the expert doesn't know that the expert favors q to at least degree t—so that learning this information might change their opinions—the asymmetric nature of 'at least degree t' plausibly means that this added information can only *favor* q further: it'll change the expert's opinions *in a predictable direction*. Thus you know that upon learning what you've learned, the expert's credence will still be at least t, so it makes sense for your credence be at least t upon learning this (Dorst 2020a, §§4–5).[23]

As Dorst (2020a) emphasizes, Simple Trust is symmetric—substituting $\neg p$ for q and $1 - s$ for t gives us the principle that conditional on the expert being *doubtful* of p, you should be doubtful of p: $\pi(p|P(p) \leq s) \leq s$. Thus Simple Trust says that when you learn whether the expert favors or disfavors q (whether $P(q) \geq t$ or $P(q) < t$), you should agree with the expert in (dis)favoring q.[24]

Dorst shows that Simple Trust is connected to Value. He first generalizes Simple Trust to apply to conditional probabilities as well as unconditional ones, yielding:

Trust: $\pi(q|p \wedge [P(q|p) \geq t]) \geq t$
Conditional on the expert being confident of q conditional on p, be confident of q conditional on p.

He then shows that Trust follows from Value (Dorst 2020a, Theorem 7.2) and that, in a natural subclass of frames, Value also follows from Trust (Theorem 7.4).[25] He further conjectures that the two are equivalent more generally.

Unfortunately, he's wrong:

Fact 2.1. *There are $\pi, \langle W, \mathcal{P}\rangle$ such that π trusts $\langle W, \mathcal{P}\rangle$ without valuing it.*

Proof. Let $\pi = \begin{pmatrix}0.17 & 0.56 & 0.27\end{pmatrix}$ and $\langle W, \mathcal{P}\rangle = \begin{pmatrix}0.45 & 0.10 & 0.45\\ 0.15 & 0.70 & 0.15\\ 0.30 & 0.10 & 0.60\end{pmatrix}$. Using the `checkTrust` function in the Mathematica notebook shows that π trusts the frame.[26] But let $O_0 = 0$ everywhere and let $O_1(w_1) = 29$, $O_1(w_2) = -3$, and $O_1(w_3) = -13$. Each $\mathcal{P}_i$ in the frame has higher expectation for O_1 than O_0, so the recommended strategy is to take O_1 everywhere. Thus $\mathbb{E}_\pi(S) = E_\pi(O_1) = -0.26 < 0 = \mathbb{E}_\pi(O_0)$; Value fails. □

Nevertheless, it turns out that Trust can naturally be generalized to yield a principle that really is equivalent to Value.

The simplest way to motivate the generalization invokes the idea of a **random variable**—a function (*any* function) from worlds to numbers. Think of this as a definite description for a number—'the number of planets'; 'the weight of this cow', 'the amount of utility you'd get from eating cake', etc. (Notice that *options*, in the sense used to specify decision problems, are simply random variables.) If you're unsure what value a random variable takes, you can form an *estimate* of it by averaging the various possibilities to try to be as close as possible. Precisely: the **expected value** of a random variable relative to a distribution π is a weighted average of the various possible values it might take, with weights determined by how likely they are to be actual.[27] Note that a proposition q is interchangeable with its **indicator variable** $\mathbb{1}_q$, i.e. the variable that assigns 1 to worlds where q is true and 0 to those where it's false. The probability assigned to a proposition equals the estimate assigned to its indicator variable: $\mathbb{E}_\pi(\mathbb{1}_q) = \pi(q)1 + \pi(\neg q)0 = \pi(q)$.

Thought of in this way, Simple Trust says that for *certain* random variables—namely, indicator variables—conditional on the expert's estimate of that variable being at least t, you have an estimate that's at least t. To formalize this, let '$\mathbb{E}(\boldsymbol{X})$' be a definite description for the expert's expectation of X (so, for example, $[\mathbb{E}(X) \geq t] = \{w : \mathbb{E}_w(X) \geq t\}$). Then Simple Trust is simply the requirement that, for every indicator variable $\mathbb{1}_q$: $\mathbb{E}_\pi(\mathbb{1}_q | \mathbb{E}(\mathbb{1}_q) \geq t) \geq t$.

It's then easy to spot a generalization. Say that π **totally trusts** a frame iff:

Total Trust: For *any* variable X: $\mathbb{E}_\pi(X \mid \mathbb{E}(X) \geq t) \geq t$
Conditional on the expert having a high estimate for X, have a high estimate for X.[28]

Total Trust is of the same form as Trust—both $[\mathbb{E}(X) \geq t]$ and $[P(q) \geq t]$ (i.e., $[\mathbb{E}(\mathbb{1}_q) \geq t]$) assert that the expert's probability function has a certain lower-bounded feature: its estimate of some quantity (X, or $\mathbb{1}_q$) is above a given threshold. So while Total Trust is stronger than Simple Trust, it formalizes a similar idea. In particular, it is likewise not affected by the criticism of Reflection we discussed earlier. While it may be that the expert doesn't know that the expert has a high estimate for X—so that this is new information that the expert has not yet taken into account—the asymmetric nature of 'high estimate' plausibly means that this added information can only *increase* their estimate for X, meaning you know that it'll still be high once this information is added.

Like Trust, Total Trust is again symmetric: π totally trusts a frame iff for all Y, s: $\mathbb{E}_\pi(Y | \mathbb{E}(Y) \leq s) \leq s$. Thus Total Trust says that upon learning whether the expert's estimate for X is high or not (whether $\mathbb{E}(X) \geq t$ or $\mathbb{E}(X) < t$), you should follow their estimate across this dividing line. (But, again, it does *not* follow that we can combine these conditions to arrive at $\mathbb{E}_\pi(X | t \leq \mathbb{E}(X) \leq t) = t$.)

Total Trust is also, in some respects, more elegant. While Trust is stronger than Simple Trust, Total Trust already implies the analogous principle for conditional estimates (endnote 36): $\mathbb{E}_\pi(X | q \wedge [\mathbb{E}(X | q) \geq t]) \geq t$; that means Total Trust implies Trust (let X be an indicator variable), and hence that Total Trust implies New Reflection as well (see endnote 37 below). Moreover, Total Trust also implies a version of itself which applies to comparisons of two estimates, rather than comparisons of an estimate with a threshold.[29]

Most importantly, Total Trust is equivalent to Value:

Theorem 2.2. *π totally trusts $\langle W, \mathcal{P} \rangle$ iff π values $\langle W, \mathcal{P} \rangle$.*[30]

Proof Sketch. $(\Rightarrow)$: Suppose $\mathbb{E}_\pi(S) < \mathbb{E}_\pi(O)$ for some $O \in \mathcal{O}$, so $\mathbb{E}_\pi(O - S) > 0$. Assume (without loss of generality, as an excruciating proof in the appendix shows) that for each $\mathcal{P}_i$ in the frame, there is a unique $O \in \mathcal{O}$ with maximal expectation. Then, finding a pair $\langle j, O_j \rangle$ that maximizes

the quantity $\mathbb{E}_j(O_j - S)$ in the frame, there will be a $t > 0$ such that $[\mathbb{E}(O_j - S) \geq t]$ includes all and only worlds w where $S_w = O_j$ and hence $(O_j - S)(w) = 0$. Thus $\mathbb{E}_\pi(O_j - S | \mathbb{E}(O_j - S) \geq t) = 0 < t$; Total Trust fails.

($\Leftarrow$): If $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) \leq t - \varepsilon$ for some X, t and $\varepsilon > 0$, then we let our options be $\mathcal{O} = \{X, Y\}$ where $Y = t - a\varepsilon$ for $0 < a < 1$ at all worlds. As $a \to 0$, we reach a point at which if S is recommended, S selects X at w if $w \in [\mathbb{E}(X) \geq t]$, and selects Y otherwise. Thus $\mathbb{E}_\pi(S)$ is an average of $\mathbb{E}_\pi(S|\mathbb{E}(X) < t) = \mathbb{E}_\pi(Y|\mathbb{E}(X) < t) = t - a\varepsilon$ and $\mathbb{E}_\pi(S|\mathbb{E}(X) \geq t) = \mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) \leq t - \varepsilon < t - a\varepsilon$, and so is less than $\mathbb{E}_\pi(Y) = t - a\varepsilon$; Value fails. □

Why does Total Trust succeed where Simple Trust and Trust failed? Inspecting the proof, clearly part of the answer is that Total Trust can apply directly to arbitrary random variables—the language of Value and its "options"—whereas Trust is restricted to propositions. But to understand why this restriction is limiting, and to identify another sense in which Total Trust is a natural generalization of Trust, it helps to visualize what these various principles require.[31] (Readers less interested in the details can skip to the next section without significant loss of continuity.)

To do this, think of a probability function π defined over n worlds as a vector in n-dimensional Euclidean space, where the ith coordinate is $\pi(w_i)$. We can then represent probability functions as points in a *barycentric plot*: a simplex (the n-dimensional generalization of a triangle) in which the extreme points are those which assign maximal probability to a single world. For a 3-world frame, this simplex is a 2D equilateral triangle. To see how this works, look at the top left triangle in Figure 4—ignore the shaded regions and arrows for now. In this figure, any point within the triangle represents a probability function. The red point in the bottom left, labeled w_1, represents the probability function that assigns 1 to w_1 and 0 to each of w_2 and w_3. (In Euclidean 3-space, this is the point $\begin{pmatrix}1 & 0 & 0\end{pmatrix}$.) Similarly, the gray point labeled w_2 at the top represents the probability function that assigns 0 to w_1 and w_3, and 1 to w_2. (The point $\begin{pmatrix}0 & 1 & 0\end{pmatrix}$.) Within the triangle, how close each dot is to these extreme points represents how confident it is in the corresponding world. For instance, $\mathcal{P}_1$ is equally confident in w_1 and w_3, hence it is in the middle of the triangle, but is much less confident of w_2, hence it is much further away from the top. (Exactly, $\mathcal{P}_1 = \begin{pmatrix}0.45 & 0.1 & 0.45\end{pmatrix}$.) In contrast, $\mathcal{P}_3$ is equally doubtful of w_2, but is slightly more confident of w_3 than of w_1, so it is the same vertical height as $\mathcal{P}_1$ but is shifted further to the right. (Exactly, $\mathcal{P}_3 = \begin{pmatrix}0.3 & 0.1 & 0.6\end{pmatrix}$.)

Using such diagrams, we can visualize what our various Trust principles require.

Start with Simple Trust. This requires that conditional on the expert being confident of q—that is, conditional on P being in the set $\{\rho : \rho(q) \geq t\}$—$\pi$ must also be confident of q (must also be in this set). In our diagrams, such "probability-threshold" sets correspond to those probability functions that fall on one side of a certain straight *cut* through the space. For example, the gray region on the top left of Figure 4 is the set of functions that assign at least $\frac{2}{3}$ to $\{w_2\}$; and the blue region on the top right is the set of functions that assign at least $\frac{2}{3}$ to $\{w_2, w_3\}$. A "cut" is a *hyperplane*—a flat surface in n-dimensional space (see below). All the probability functions that are in the probability-threshold set are those that fall above this hyperplane.

Simple Trust says that conditional on the expert being in this set, you should be in this set. Since Simple Trust is symmetric, it also says that conditional on the expert being *outside* this set (in the non-shaded region), you should be outside this set. What Simple Trust amounts to is thus: *trust the expert's judgment when you learn which side of a probability-threshold cut they're on*. As can be seen, π in the top row of Figure 4 satisfies this criterion for the two cuts we displayed: the solid arrows indicate where π moves when it conditions on P being in the shaded set ($P(q) \geq t$); the dotted arrow indicates where it moves when it conditions on P being in the non-shaded set

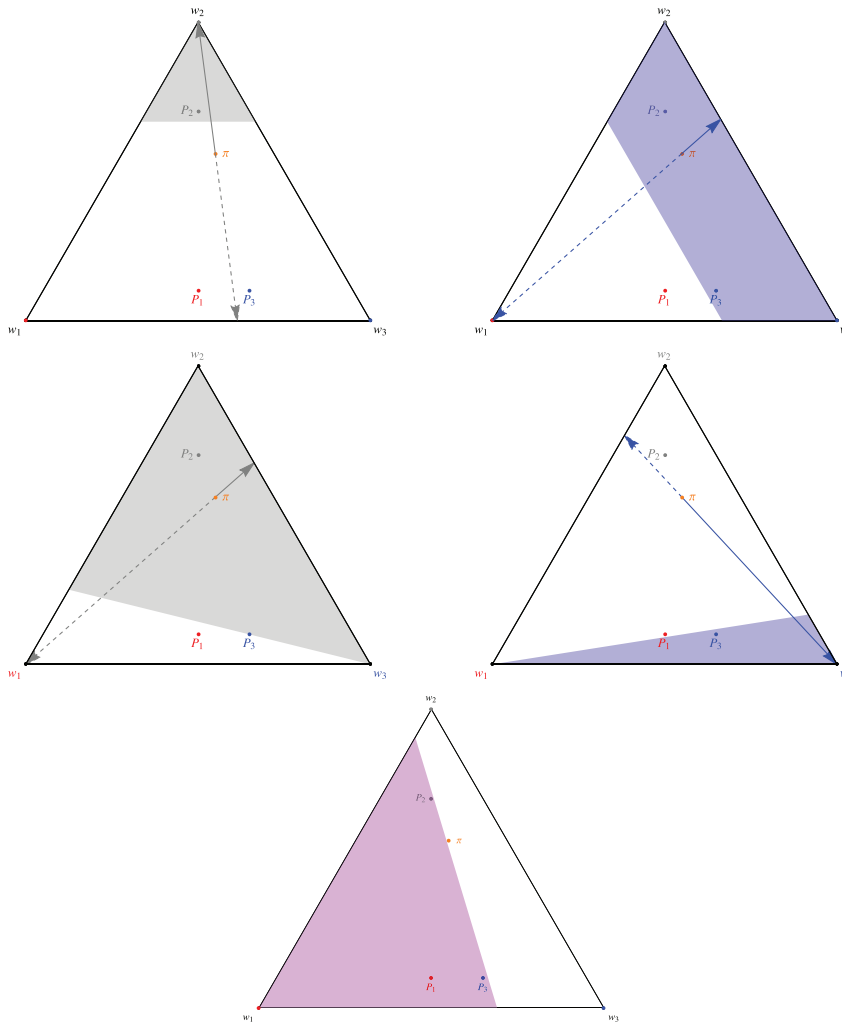


FIGURE 4 A frame and probability function π displayed with several different cuts. **Top**: Two probability-threshold sets governed by Simple Trust. Gray region (left figure) comprises $\{\rho : \rho(w_2) \geq \frac{2}{3}\}$; blue region (right figure) comprises $\{\rho : \rho(\{w_2, w_3\}) \geq \frac{2}{3}\}$. Solid arrows represent where π moves when it conditions on P being in the shaded set; dotted arrows are where it move when it conditions on P being in the non-shaded set. **Middle**: Two conditional-probability-threshold sets governed by Trust. Gray region (left figure) comprises $\{\rho : \rho(w_2|\{w_1, w_2\}) \geq {}^1/_4\}$; blue region (right figure) comprises $\{\rho : \rho(w_3|\{w_2, w_3\}) \geq {}^5/_6\}$. Solid arrow represents where π moves when it conditions on P being in the shaded set; dotted arrow represents where it moves when it conditions on P not being in the shaded set. **Bottom**: A cut governed by Total Trust. π does not trust the expert's judgment across this cut—conditional on the expert being to the left of the cut, π stays to the right, where it already is—meaning that Total Trust fails

$(P(q) < t)$. For instance, the solid gray arrow in the top left maps π to w_2 (makes π certain of w_2) since w_2 is the only world w such that $\mathcal{P}_w(w_2) \geq \frac{2}{3}$—as can be seen by the fact that $\mathcal{P}_2$ is the only world in the gray region. Now turn from Simple Trust to Trust. This generalization matters because it introduces additional cuts. In our simple three-world diagrams, the cuts associated with probability thresholds are all and only those that are parallel to one of the sides of the triangle. By

strengthening Simple Trust to Trust, we require that you trust the expert's judgment about certain additional cuts—namely those corresponding to *conditional* probability thresholds. In our three-world diagrams, this adds all the cuts that intersect one of the vertices—see the middle row of Figure 4 for two examples. Though it is tricky to verify in full generality, it turns out that π trusts P's judgment about *all* such cuts, i.e. π trusts this frame.

Turn, finally, to Total Trust. The key realization is that there are *further* cuts beyond those governed by Trust: not every cut corresponds to a conditional-probability threshold. In particular, consider the purple region in the bottom row of Figure 4. This region is bounded by a cut; and since this cut separates π and the $\mathcal{P}_i$, π clearly does not trust the expert's judgment across this cut. After all, π is already certain that the expert is on the other side; so conditional on the expert being on that side, π remains exactly where it is.

Now for the big reveal: the case we've been diagramming is in fact the case that we used above (Fact 2.1) to show that π can trust a frame without valuing it. So the fact that π agrees with P across all of the cuts corresponding to probability and conditional-probability thresholds just is the fact that π trusts this frame. And the purple line separating π from the frame represents the set of probability functions ρ such that $\mathbb{E}_\rho(O_1) = 0$—where O_1 is as defined in the decision problem used to show that π doesn't value the frame, and 0 is the fixed utility of the alternative option O_0. So the failure of π to value the frame corresponds to its failure to trust the expert about these additional cuts; requiring π to trust the expert about *every* cut would clearly be sufficient for eliminating this example.[32]

There is, in fact, a more general connection between "cuts" and random variables such as O_1. For a **cut** is a hyperplane, and a hyperplane is specified by a linear equation; so a "cut" is a set $\{\rho : \rho(w_1)x_1 + \cdots + \rho(w_n)x_n = t\}$, for some $x_1, \ldots, x_n, t \in \mathbb{R}$. The two sides into which this cut separates probability-space are the sets $\{\rho : \rho(w_1)x_1 + \cdots + \rho(w_n)x_n \geq t\}$ and $\{\rho : \rho(w_1)x_1 + \cdots + \rho(w_n)x_n \leq t\}$. But now we can think of $x_1, x_2, \ldots, x_n$ as the values of some random variable X, meaning that a cut is simply a set $\{\rho : \mathbb{E}_\rho(X) = t\}$ for some t and X, and the two sides are $\{\rho : \mathbb{E}_\rho(X) \geq t\}$ and $\{\rho : \mathbb{E}_\rho(X) \leq t\}$. The requirement to trust the expert's judgment about *every* cut is thus exactly the requirement that, for every random variable X and threshold t, your expectation for X conditional on the expert's expectation being greater than t should also be greater than t (and *mutatis mutandis* for *smaller than t*). That is, it is simply the principle Total Trust. *That* is why Total Trust succeeds where Simple Trust and Trust failed.

This way of thinking about our Trust principles may suggest ideas for even stronger ones. After all, why should the principle be restricted to cuts, i.e. hyperplanes, i.e. *flat* surfaces. Wouldn't it be natural to also require π to trust the expert's judgment about *bent* surfaces? Or to go further yet and simply require that, for any condition C on probability distributions, π should agree with the expert on whether to exhibit C when it learns whether the expert does so—that is, $\pi(\cdot|\mathcal{P} \in C) \in C$ and $\pi(\cdot|\mathcal{P} \notin C) \notin C$?

But any such strengthened requirement would be too demanding for reasons quite orthogonal to anything arising from modesty. Suppose that there are two candidate expert distributions, ρ_1 and ρ_2, with extreme probabilities: that $\rho_1(q) = 1$ and $\rho_2(q) = 0$. π can clearly defer to this expert even if, upon learning that the expert's credences are either ρ_1 or ρ_2, π assigns a probability between 0 and 1 to q—conditional on the expert knowing whether q, you may still be uncertain whether q! (Even Reflection allows this.) But assigning q a probability between 0 and 1 conditional on P being either ρ_1 or ρ_2 means that π does not agree with the expert on whether to exhibit the condition $\{\rho : \rho(q) = 1 \text{ or } \rho(q) = 0\}$. More generally, say that a set C of probability distributions is **convex** if it contains every distribution that is a weighted average of others it contains.[33] It seems that you can defer to an expert by adopting any weighted average of the possible opinions of the

expert. (This, too, is allowed by Reflection.) So if C is not convex, π can defer to $\mathcal{P}$ even though though $\pi(\cdot|P \in C) \notin C$. So at most it makes sense to require $\pi(\cdot|\mathcal{P} \in C) \in C$ if C is convex.

But notice: hyperplanes are the *only* divisions that divide probability space into two convex sets.[34] So if we want a principle that respects the basic shape of Total Trust—requiring you to trust the expert's judgment on either side of a division—that principle should be no stronger than Total Trust itself.[35] This observation helps situate the relative strengths of Reflection, Total Trust, and Simple Trust. In particular, Reflection can be equivalently stated using convex sets:

Reflection (convexity version): $\pi(\cdot|P \in C) \in C$, for any convex C.

Likewise for Total Trust. Say that a set B is **biconvex** iff both B and its complement are convex. Then since a set is biconvex iff its boundary is a hyperplane (endnote 34), Total Trust is equivalent to:

Total Trust (convexity version): $\pi(\cdot|P \in B) \in B$, for any *bi*convex B.

This formulation of Total Trust is perhaps the most formally useful; it allows us to easily show that Total Trust implies a variety of other principles. For instance, Total Trust implies a seemingly more general conditional-expectation version of our above formulation.[36] This shows that Total Trust implies Trust (let X be an indicator variable), which in turn shows that Total Trust implies New Reflection.[37] Likewise, Total Trust implies that conditional on the expert's *average* credence across a set of propositions $\{q_i\}$ being at least t, your average credence should also be at least t.[38]

Finally, this convexity formulation allows us see why Simple Trust remains a natural component of Total Trust. Suppose we restrict ourselves to *proposition-level* formulations of such deference principles—that is, ones that use only conditions on P that can be stated in terms of the expert's opinion about a single proposition q. The set of *convex* proposition-level conditions are those of the form $P \in \{\rho : \rho(q) \in [l, h]\}$, i.e. those that say that the expert's credence in a given q is within a given range $[l, h]$. This leads to a proposition-level formulation that is only very slightly weaker than Reflection as formulated above: $\pi(q|P(q) \in [l, h]) \in [l, h]$ (see Gallow 2017). Meanwhile, the set of *biconvex* proposition-level conditions are those of the form $P \in \{\rho : \rho(q) \geq t\}$, i.e. those that say the expert's credence in a given q is above a given threshold t. That is, restricting Total Trust to proposition-level conditions gives us exactly Simple Trust: $\pi(q|P(q) \geq t) \geq t$. Thus Simple Trust is the strongest component of Total Trust (i.e. Value) that we can formulate if we restrict ourselves to deference principles built around single propositions. That restriction is where Simple Trust went wrong.

A similar lesson applies to New Reflection. If we restrict ourselves to learning about the *full* distribution of P—i.e. propositions of the form $[P = \rho]$—then New Reflection is the strongest thing we can say. More precisely, New Reflection is the strongest principle that follows from Total Trust (i.e. Value) of the form $\pi(\cdot|P = \rho) \in C$, for some C.

The obvious lesson from these two instances is that the *form* of potential deference principles we countenance can easily impose problematic restrictions in our search for the correct one. It is only when a range of different forms coincide—as they do with Value, the various formulations of Total Trust, and several other formulations we'll see below—that we can have confidence in the principle we end up with.

3 | THE VALUE OF ACCURACY

We've now shown that Total Trust captures the pragmatic version of Value we've been focusing on: the principle that, no matter what decision you face, you should always prefer to let the expert decide for you rather than decide yourself. But there is also an *epistemic* version of this constraint—namely, that you should always expect the expert to *be more accurate* than you are. In other words, if all you care about is getting to the truth, you should prefer to have the expert's credences, rather than your own:

Epistemic Value (rough): Always expect the expert's opinions to be more accurate than your own, under any reasonable way of measuring accuracy.

It's straightforward to see that Value entails Epistemic Value. After all, here's a decision-problem: adopt a credence function, and receive utility in proportion to its accuracy. Standardly, every candidate for the expert (every probability function) must expect itself to be more accurate than any particular alternative candidate—this is the *strict propriety* constraint on accuracy measures required for virtually all the major results in the literature.[39] This entails that the recommended strategy S in such a decision problem is always simply to adopt the expert's credences, whatever they are. Since Value entails that this strategy has higher expected utility than simply sticking with your credence function, it entails that you expect the expert to be more accurate than you.[40]

The converse, however, is not straightforward: why should expecting the expert to be more accurate than you (have preferable credences for *one* type of decision—namely, what opinions to have) necessarily mean that you prefer to use their opinions for *all* possible decisions? A natural first thought is that the *reason* you prefer to use the expert's opinions to make decisions is that you expect them to be more accurate than yours. But, as we'll see, it turns out that the tenability of this thought depends heavily on how wide the range of reasonable ways of measuring accuracy is.

In particular, the recent work of Levinstein (2019) can be marshaled to show both a close connection but also a potential divergence between Value and Epistemic Value. There is a large literature devoted to measures of accuracy ('scoring rules') and the constraints they can be used to impose on rational opinions.[41] By filling in 'any reasonable way of measuring accuracy' in Epistemic Value with the standard class of scoring rules used in this literature, Levinstein (2019) shows that this version of Epistemic Value is equivalent to *Simple* Trust.

Since we now know that Simple Trust is substantially weaker than Value, this raises the possibility that Epistemic Value and Value might come apart. But as we'll show, we need not accept that conclusion. The results of Campbell-Moore (2020) suggests that we can broaden our view of what counts as a 'reasonable way of measuring accuracy' by broadening what *sort* of state we can measure the accuracy of: instead of just measuring the accuracy of credences in propositions, we can measure the accuracy of estimates of random variables. Once we broaden our view in this way, we'll show that Epistemic Value does indeed turn out to be equivalent to Value. (Those uninterested in the details can skip to §3.1.)

First, we need to get clear on what Levinstein (2019) shows. Let I_q be a local *in*accuracy measure for a given proposition q: it takes a probability function, δ, and truth-value of q, $\mathbb{1}_q$ and outputs a non-negative real number measuring the divergence between the probability and the truth of q—i.e. how inaccurate δ is about q. $I_q(\delta)$ can then be treated as a random variable—'the inaccuracy of δ about q, whatever it is.' Say that I_q is *truth-directed* iff being closer to the truth-value of q

makes a probability function more accurate (less inaccurate).[42] Say that I_q is *strictly proper* iff every probability function expects itself to be more accurate than any other (rigidly designated) probability function.[43]

Now let I be a *global* inaccuracy measure which takes a probability function δ and a world w and outputs its overall inaccuracy at w. Say I is *additive* iff it is a sum of local inaccuracy measures.[44] Say that an additive I is truth-directed and strictly proper iff all its component local scoring rules are truth-directed and strictly proper. The favored class of scoring rules within the epistemic utility literature is the class of additive, truth-directed, strictly proper scoring rules (Predd et al. 2009; Pettigrew 2016; Levinstein 2017a; Campbell-Moore and Levinstein 2020). So suppose we assume that *these* are all and only the reasonable ways of measuring accuracy; then Epistemic Value corresponds to Simple Trust.

Notice that $I(P)$ and $I_q(P)$ can be treated as random variables for 'the inaccuracy of the expert's opinions, whatever they are': $I_q(P)(w) = I_q(\mathcal{P}_w, w)$, etc. Now fixing some particular q, say that π simply trusts P *with respect to* q iff for all t: $\pi(q|P(q) \geq t) \geq t$ and $\pi(q|P(q) \leq t) \leq t$. Then we have:

Theorem 3.1 (Levinstein 2019). *π simply trusts P with respect to q iff for every continuous,[45] truth-directed, strictly proper local scoring rule I_q, $\mathbb{E}_\pi(I_q(P)) \leq \mathbb{E}_\pi(I_q(\pi))$, with equality if and only if $\pi(P(q) = \pi(q)) = 1$.*

A corollary is that π simply trusts P (for *all* propositions) if and only if for every continuous, truth-directed, strictly proper, and additive *global* scoring rule I, $\mathbb{E}_\pi(I(P)) \leq \mathbb{E}_\pi(I(\pi))$.

What to make of this result? If this is the correct class of 'reasonable ways of measuring accuracy', then—since Simple Trust is strictly weaker than Total Trust—it follows that Epistemic Value is strictly weaker than Value.

But an alternative reading is possible: the result may suggest that the standard way of thinking about scoring rules is overly narrow; they do *not* capture every reasonable way of measuring accuracy. To see why, return to the example used in Fact 2.1 and the bottom row of Figure 4. As we've seen, in that example π trusts P, and therefore expects P to be more accurate than itself on all of the ways of measuring accuracy that Theorem 3.1 countenances. Yet there is an issue on which, it seems, π may sensibly *not* expect P to be more accurate than itself on—namely, the value of the random variable O_1. π estimates O_1 to have a value of -0.26, while every candidate expert function estimates it to have a value of at least 0.3—this was why we were able to separate π from the frame using the purple cut in the bottom row of Figure 4.

Visually, there's an intuitive sense in which π could expect itself to be more accurate about O_1 than the expert—namely, it's on the correct side of the line dividing the purple from white region. For example, suppose we measure accuracy this way: to have an accurate estimate of O_1, it matters a lot whether your estimate is on the correct side of 0, but very little how close it is beyond that. On that way of measuring accuracy of estimates, π will expect itself to be more accurate about O_1 than the expert.[46]

This suggests that we should follow the work of Campbell-Moore (2020) and consider scoring rules that apply to arbitrary estimates—not just to estimates of indicator random variables, i.e. probabilities of propositions. If we do so, we'll find that Value and Epistemic value will align exactly.

Precisely, let an **estimate-inaccuracy measure** for a random variable X take a estimate $e \in \mathbb{R}$, a world w, and output the inaccuracy of e at w, denoted $I_X(e, w)$. Writing $I_X(\pi)$ to abbreviate $I_X(\mathbb{E}_\pi(X))$, say that I_X is **generally strictly proper (gsp)** iff any probability function expects its

own estimate of X to be more accurate than any other (rigidly designated) estimate.[47] For tractability, assume I_X is absolutely continuous in its first argument.

In a recent paper, Campbell-Moore (2020) has shown that Schervish's (1989) well-known characterization of strictly proper local inaccuracy measures can be generalized to characterize gsp estimate-inaccuracy measures as well. Using this more general characterization, we can show that our pragmatic version of value *does* coincide with a version of Epistemic Value—namely, the version we get if we say that the class of 'reasonable ways of measuring (in)accuracy' correspond exactly to the generally strictly proper *estimate*-scoring rules. In fact, we can show an even tighter connection than that—one that outstrips in *both* directions the argument that Value implies Epistemic Value from above. For we can show that totally trusting an expert *with respect to a given variable* X is equivalent to expecting their estimate of X to be more accurate than your own under an way of measuring the accuracy of such an estimate.

Precisely, let $I_X(P)$ be the inaccuracy of the expert's estimate for X, whatever it is: $I_X(P, w) = I_X(\mathbb{E}_w(X), w)$. Say that π **epistemically values** the P *with respect to* X iff it expects P's estimate of X to be more accurate than its own, on any (generally strictly proper) way of measuring accuracy.[48] And say that π totally trusts P *with respect to* X iff for all t: $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) \geq t$ and $\mathbb{E}_\pi(X|\mathbb{E}(X) \leq t) \leq t$. Then:

Theorem 3.2. *π totally trusts P with respect to X iff π epistemically values P with respect to X.*[49]

Proof Sketch. ($\Rightarrow$): Let $\mathbb{E}_\pi(X) = e$. Given Total Trust, we show:

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) > e) < \mathbb{E}_\pi(I_X(e) \mid \mathbb{E}(X) > e) \tag{1}$$

This suffices for the proof since a symmetric argument shows:

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) < e) < \mathbb{E}_\pi(I_X(e) \mid \mathbb{E}(X) < e) \tag{2}$$

Jointly equations (1) and (2) entail this direction of the theorem.

To prove equation (1): Let $w_1, \ldots, w_p$ be the worlds where for each i, $\mathbb{E}_i(X) > e$. Without loss of generality, assume that for each $i < p$, $\mathbb{E}_i(X) > \mathbb{E}_{i+1}(X)$. We then prove by induction for all k with $1 \leq k \leq p$ and for any $s < \mathbb{E}_k(X)$:

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) \geq \mathbb{E}_k(X)) < \mathbb{E}_\pi(I_X(s) \mid \mathbb{E}(X) \geq \mathbb{E}_k(X))$$

($\Leftarrow$): Suppose that $\mathbb{E}_\pi(X \mid \mathbb{E}(X) \geq t) < t$ for some t. (The case where $\mathbb{E}_\pi(X \mid \mathbb{E}(X) < t) \geq t$ can be treated similarly.) Then there is some region (α, β) where for all $t \in (\alpha, \beta)$, $\mathbb{E}_\pi(X \mid \mathbb{E}(X) \geq t) < t$. In Appendix B.3.1, we show how to make a gsp that pays special attention (α, β). It is then relatively straightforward to show that $\mathbb{E}_\pi(I_X(e)) < \mathbb{E}_\pi(I_X(P))$ for this rule. □

It follows immediately from Theorems 2.2 and 3.2 that Value is equivalent to Epistemic Value—at least, it is if we understand the 'reasonable ways of measuring accuracy' as the set of *all* gsp estimate-inaccuracy measures:

Epistemic Value: For any X and gsp I_X, $\mathbb{E}_\pi(I_X(P)) \leq \mathbb{E}_\pi(I_X(\pi))$ with equality if and only if $\pi(\mathbb{E}(X) = \mathbb{E}_\pi(X)) = 1$.
Always expect the expert's estimates to be more accurate than your own.

As with Theorem 3.1, this result immediately generalizes to additive, "global" estimate-inaccuracy which measure the inaccuracy of a given probability function by a sum of the inaccuracies of its various estimates.

3.1 | Upshots

By Theorem 3.2, you expect an expert's estimate of a particular quantity X to be more accurate than your own on every reasonable way of measuring accuracy iff you totally trust their estimate of that quantity. This result carries with it some philosophical subtleties, so let's pause to take a look.

First subtlety: recall that many potential experts are only *local* experts—you should only trust them on a restricted range of questions (Introduction). Unlike Theorems 2.2 and 4.1, Theorem 3.2 is a *local* equivalence result. Fix some particular quantity X—say, the number of inches of rain we'll have next week. And fix on some potential expert—say, the weather forecaster. The result says that you totally trust the forecaster *about how much rain we'll have* iff you expect their estimate to be more accurate than your own on every reasonable way of measuring its accuracy. This holds even if you don't totally trust them on other questions (like what to do with the now-far-too-old bananas). This is significant because it shows that the connection between Total Trust and Epistemic Value is a very tight one. In particular, the connection is not restricted to modest experts. As we'll discuss below (§5), you can totally trust an *im*modest expert with respect to a certain question (or with respect to certain quantities) even without reflecting them with respect to that question. By Theorem 3.2, you will then likewise expect them to be more accurate than you with respect to that question. Thus this theorem shows that even when it comes to immodest experts, Total Trust carves out a formally and philosophically natural notion of deference that is weaker than Reflection.

Second subtlety: note that the biconditional connecting Total Trust to Epistemic Value goes through only on a very permissive account of the reasonable ways of measuring accuracy. If there are *more* constraints on reasonable ways of measuring accuracy beyond them being gsp estimate-inaccuracy measures, then we would lose the right-to-left direction of the proof, and Epistemic Value might not entail Value. Yet some have argued that there *are* further constraints—for example, that the Brier score (squared Euclidean distance) is the uniquely reasonable scoring rule (Joyce 2009; Pettigrew 2015, 2016). What should we make of this discrepancy?

One possibility is to endorse a form of subjectivism about measuring accuracy: individuals have complete latitude (amongst the gsp estimate-inaccuracy measures) to choose how they are going to value accuracy. The proper formulation of Epistemic Value should then guarantee that if they defer to an expert, then *however* they decide to measure accuracy, they will expect the expert to be more accurate than their own credence function. A second reaction would be a form of objectivism, combined with an epistemic-robustness constraint: although there is a uniquely rational way of measuring accuracy, reasonable people can be uncertain what it is (amongst the gsp estimate-inaccuracy measures). The proper formulation of Epistemic Value should then guarantee that no matter how this uncertainty is distributed, if they defer to an expert then they'll expect them to be more accurate than their own credence function. A final option is a form of supervaluationism: it's indeterminate (amongst the gsp estimate-inaccuracy measures) what the correct way of measuring accuracy is, but if you defer to an expert then it's determinately true that you expect them to be more accurate than you—therefore, they must be more accurate on all such measures.

We are neutral between these three approaches. But we do think that Theorem 3.2 lends some support to a reading along one of these lines—i.e. to a pluralistic approach to accuracy-measures. For it turns out that the universal quantification over accuracy measures is extremely important in the above argument: merely requiring that π expects P to be more accurate than itself *according to (say) the propositional Brier score*, does not entail any interesting deference principle—not even Simple Trust or New Reflection.[50] Interestingly, requiring π to expect *all* of P's estimates to be more accurate than its own as measured by squared Euclidean distance *may* entail New Reflection (we've been unable to find a counterexample). But New Reflection does not entail this constraint, and the constraint still does not entail even Simple Trust.[51]

Thus insofar as there's reason to want a tight connection between deference and expected accuracy, there's reason to be pluralist about the acceptable ways of measuring accuracy. And, we think, there *is* such reason: after all, it's intuitive to think that when you value an expert, the *explanation* for why you prefer to use their opinions to make decisions is that you expect their opinions to be more accurate than your own. As we've seen, this intuition is correct iff the reasonable ways to measure accuracy correspond to the set of gsp estimate-inaccuracy measures.

4 | THE GEOMETRY OF DESERVED DEFERENCE

So far we've shown how to generalize Trust to arrive at a deference principle—Total Trust—that can be stated as a constraint on the relationship between your opinions and the expert's, and which characterizes what it takes to value an expert's opinions for the sake of making good decisions or accurate estimates. But this doesn't yet tell us exactly what constraints the various candidates for the expert must meet in order to deserve π's deference. In particular, we'd like a characterization of these constraints along with the relationship π must bear to them such that, if we are given a probability function and a frame, we can (efficiently) check whether the function values that frame. (Compare: in epistemic logic, we don't simply want to know which axioms are equivalent the KK principle that $Kq \to KKq$; we also want to know that a frame validates this principle iff it is transitive.) This section will give such a characterization, revealing another way in which Total Trust is a natural middle ground between Reflection and New Reflection. (This will require attention to some technicalities; if you are just here for the philosophy, you may prefer to skip to §5.)

The characterization begins, once more, by looking at pictures. Compare frames F_1 and F_2, described and pictured in the first two rows of Figure 5. The two frames look very similar: in both, each $\mathcal{P}_i$ is more confident of w_i than of the other worlds; in both $\mathcal{P}_2$ thinks w_1 is more likely than w_3; in the second, it just does so to a slightly greater extent. But while there are probability distributions that totally trust the first frame (for example, the uniform distribution $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, or any other distribution in the triangle delineated by the $\mathcal{P}_i$), there are no probability distributions that totally trust the second. Why the difference?

Notice that in F_2 we can separate $\mathcal{P}_2$ from the rest of the frame using a cut that does not include w_2 (right side of second picture-row of Figure 5). As a result, conditional on the expert being to the left of this cut, π moves directly to w_2, and does *not* trust the expert on this cut.[52] In contrast, we cannot do the same in F_1: any cut that is shallow enough to include only $\mathcal{P}_2$ will also include w_2, so π will map to w_2 and trust P across about this cut (middle picture row of Figure 5; gray region and solid arrow on the left); and any cut steep enough to exclude w_2 will include $\mathcal{P}_1$, and

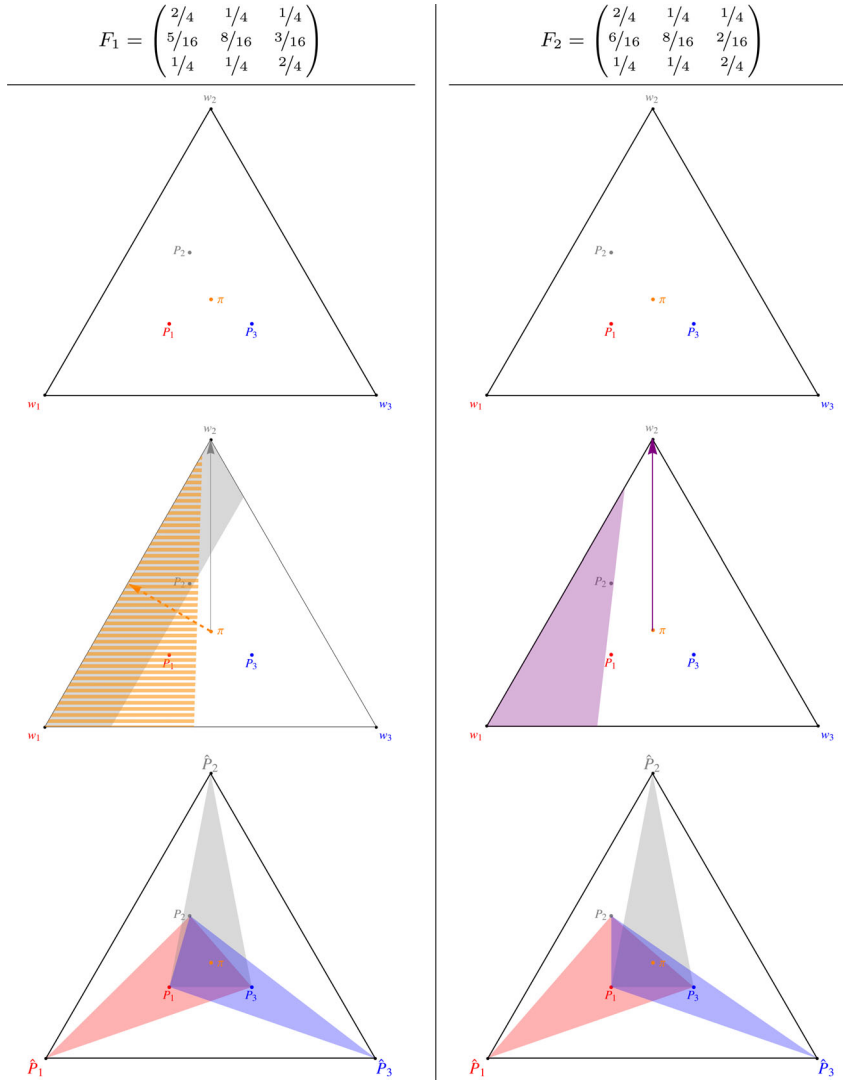


FIGURE 5 Two frames. $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ values F_1 (left column) but not F_2 (right column). **Top**: The frames are the same except that in F_2, $\mathcal{P}_2$ is slightly further weighted towards w_1 over w_3. **Middle**: In F_2 we can use a cut to separate $\mathcal{P}_2$ from its world (w_2) and the other $\mathcal{P}_i$ (right); in F_1 we cannot (left): any cut shallow enough to exclude $\mathcal{P}_1$ (gray region) will include w_2, so will map π to w_2 (solid arrow); any cut steep enough to include $\mathcal{P}_1$ (hatched region) will include the $(\frac{1}{2}, \frac{1}{2}, 0)$ point it maps π to (dotted arrow). **Bottom**: Shaded triangles represent the convex hulls of $\{\widehat{\mathcal{P}}_1, \mathcal{P}_2, \mathcal{P}_3\}$ (red/left), $\{\mathcal{P}_1, \widehat{\mathcal{P}}_2, \mathcal{P}_3\}$ (gray/top), $\{\mathcal{P}_1, \mathcal{P}_2, \widehat{\mathcal{P}}_3\}$ (blue/right). Note that in F_1, $\mathcal{P}_2$ falls within the gray hull, while in F_2 it does not.

so map π to the left side of the triangle, again trusting the expert about the cut (orange hatched region and dashed arrow).

What makes for the difference? The answer is easy to visualize. First, recall that $\widehat{\mathcal{P}}_w := \mathcal{P}_w(\cdot|P = \mathcal{P}_w)$ is the expert's *informed* opinions—those they'd have upon learning that they are the expert (§1). Since in F_1 and F_2 each $\mathcal{P}_w$ is unique, this means that $\widehat{\mathcal{P}}_w$ is certain it's at w, so is represented by the extreme point in the triangle that's certain it's at w: $\widehat{\mathcal{P}}_1$ is at the bottom left

corner, $\widehat{\mathcal{P}}_2$ is the top corner, etc. Now notice the differing relationship between the $\mathcal{P}_i$ and the $\widehat{\mathcal{P}}_i$ in F_1 and F_2. The **convex hull** of a set of points $\rho_1, \ldots, \rho_n$ is the smallest convex set containing them all, i.e. the set of points obtainable by averaging them.[53] On the bottom picture-row of Figure 5 we've plotted the convex hulls (shaded triangles) of $\{\widehat{\mathcal{P}}_1, \mathcal{P}_2, \mathcal{P}_3\}$ (red/left), $\{\mathcal{P}_1, \widehat{\mathcal{P}}_2, \mathcal{P}_3\}$ (gray/top), and $\{\mathcal{P}_1, \mathcal{P}_2, \widehat{\mathcal{P}}_3\}$ (blue/right). Once we do so, the difference betwen F_1 and F_2 jumps out: in the former, $\mathcal{P}_2$ falls insides the gray convex hull of $\{\mathcal{P}_1, \widehat{\mathcal{P}}_2, \mathcal{P}_3\}$, while in the latter it does not. *This* is what allowed us to separate $\mathcal{P}_2$ from the rest of the frame and w_2, generating a Value failure for F_2.

It also points us to a characterization. Say that a candidate $\mathcal{P}_i$ is **modestly informed** iff their opinions are an average of their own informed opinions $\widehat{\mathcal{P}}_i$ along with the (uninformed) opinions $\mathcal{P}_j$ of the other candidates they think might be an expert.[54] This is the key constraint. π totally trusts (i.e. values) a frame iff all the candidates for the expert are modestly informed, and π is an average of them:

Theorem 4.1. *π totally trusts $\langle W, \mathcal{P}\rangle$ iff each $\mathcal{P}_i \in C_\pi$ is modestly informed and π is in their convex hull.*

Proof Sketch. ($\Rightarrow$:) If π is not in the convex hull of the $\mathcal{P}_i$, we can separate it from them using a cut as we did in the bottom row of Figure 4, in which case Total Trust fails. And if one of the $\mathcal{P}_i$ is not modestly informed, we can use a cut to separate it from its worlds and the other $\mathcal{P}_j$ like we did in the bottom of Figure 5, leading to a Total Trust failure.

($\Leftarrow$:) If Total Trust fails, then Value fails, so $\mathbb{E}_\pi(O - S) > 0$ for some $\mathcal{O}$, $O \in \mathcal{O}$, and recommended S. There must be some extreme point $\mathcal{P}_i$ in the frame that maximizes this divergence. The trouble is, if $\mathcal{P}_i$ is modestly informed then $\mathbb{E}_i(O - S)$ is an average of $\widehat{\mathbb{E}}_i(O - S)$ (which is 0, since O maximizes $\mathbb{E}_i$ and so $S_i = O$) and the other $\mathbb{E}_j(O - S)$ (which are less extreme than $\mathbb{E}_i(O - S)$). This contradicts the assumption that $\mathcal{P}_i$ is an extreme point, so it must not be modestly informed. □

How should we understand this result—and, in particular, the constraint that each candidate be modestly informed?

First, to get some intuition for the constraint, consider what it amounts to in one standard case of higher-order evidence: you and some peers do some reasoning and each come to have a certain "hunch"; but then you realize your hunches differ, conclude that you might not have reasoned properly, and so adjust your opinions to take account of your higher-order doubts. For each i, we can think of the informed opinion $\widehat{\mathcal{P}}_i$ as the "hunch" of the person who reasoned properly at world i[55]—after all, they're the opinions that person would have if they had no higher-order doubts about their reasoning. Meanwhile, $\mathcal{P}_i$ represents the "all-doubts-considered" opinions of the person who reasoned properly at world i—those which that person has once they've taken their higher-order doubts into consideration. To require the rational person's (all-doubts considered) opinions to be modestly informed is thus to insist that they are some kind of average of the hunch of the well-reasoning individual, $\widehat{\mathcal{P}}_i$, and the all-doubts-considered opinions of those who reasoned poorly. It thus permits both an extreme "right reasons" response (which maintains that the all-doubts considered opinions of the well-reasoning individual should simply be their hunch; Titelbaum 2015) and an extreme "conciliationist" response (which maintains that everyone's all-doubts-considered opinions should coincide; Elga 2007), and everything that falls strictly between these.

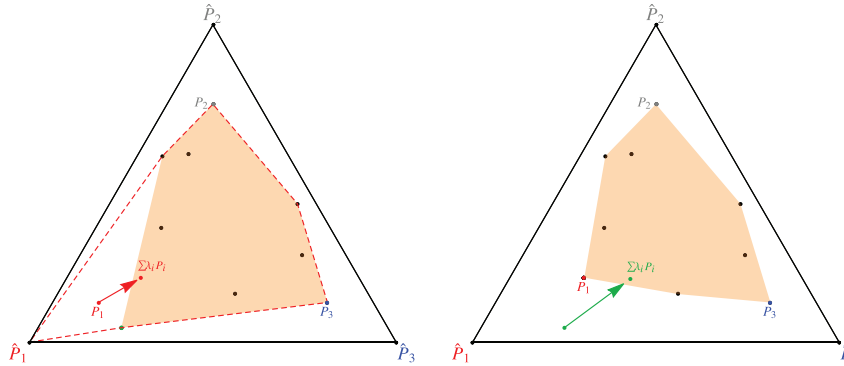


FIGURE 6 Projection onto 2D of a set of points illustrating why each $\mathcal{P}_i$ can't stably be a a nontrivial average of itself and C_i^- unless it is an average of C_i^-. **Left**: Orange region is convex hull of C_1^-; since $\mathcal{P}_1$ falls outside it, it is more extreme than C_1^-; thus averaging $\mathcal{P}_1$ with this pulls it towards the others. (In contrast, the red dotted line delineates the convex hull of $\{\widehat{\mathcal{P}}_1\} \cup C_1^-$, meaning that $\mathcal{P}_1$ *can* stably be an average of these points—i.e. it can stably be modestly informed.) **Right**: The process iterates as new extreme points are pulled towards the others, illustrating why Reflection makes modesty unstable

Now let's turn to seeing how the modestly-informed constraint relates to those imposed by Reflection and New Reflection. A helpful way to do so is to focus on what our various principles require of the expert candidates, assuming that they all defer to the expert.[56] Say that a frame **validates** a deference principle Φ iff every $\mathcal{P}_i$ in the frame defers to it Φ-wise. Reflection then is equivalent to the requirement that each $\mathcal{P}_i$ is either immodest or an average of the other candidates it leaves open:

Fact 4.2. *$\langle W, \mathcal{P}\rangle$ validates Reflection iff for each $i \in W$: either $\mathcal{P}_i$ is immodest or $\mathcal{P}_i$ is in the convex hull of C_i^-.*[57]

Why is this true? In particular, why does Reflection require that if $\mathcal{P}_i$ is modest, then $\mathcal{P}_i$ is an average of the *other* candidates, excluding itself? Reflection immediately implies that $\mathcal{P}_i$ can be written as an average of itself along with the other candidates, C_i^-.[58] Moreover, if $\mathcal{P}_i$ is modest, then $\mathcal{P}_i(P = \mathcal{P}_j) > 0$ for some $\mathcal{P}_j \in C_i^-$; so then this average is nontrivial in the sense that it gives at least some weight to the other candidates C_i^-. Now, in order for a point to fall outside the convex hull of the other candidates, it must be more extreme than of them in some direction (Figure 6). It's then easy to see that any nontrivial average of such a point and the other candidates would be less extreme than the point itself. (If $x > y_1, \ldots, y_n$, then any average of x with the y_i will be less than x.) Since $\mathcal{P}_i$ is such a nontrivial average, and can't be less extreme than itself, it follows that $\mathcal{P}_i$ must already be in the convex hull of the other candidates C_i^-.

Moreover, it is this feature which makes Reflection incompatible with modest experts. For consider the set of all the candidate experts. By the reasoning above, none of its extreme points can be modest. So all of its extreme points are immodest, and thus assign probability 0 to any candidate other than themselves. In particular, they all assign probability 0 to every modest candidate. And since any other candidate must be in their convex hull (otherwise *it* would be an extreme point), it follows that any other candidate also assigns probability 0 to every modest candidate; so modest candidates might as well not be included in the frame. The problem, in essence, is that Reflection

requires $\mathcal{P}_i$ to be an average of *itself* and the other candidates it leaves open; but a point $\mathcal{P}_i$ can't be used to anchor *itself* outside of the hull of the other candidates it gives weight to.

New Reflection solves this problem by having each $\mathcal{P}_i$ be an average of a *different* set of points—the informed $\widehat{\mathcal{P}}_j$—rather than the $\mathcal{P}_j$ themselves:

Fact 4.3. *$\langle W, \mathcal{P}\rangle$ validates New Reflection iff for each $i \in W$: $\mathcal{P}_i$ is in the convex hull of $\{\widehat{\mathcal{P}}_j : \mathcal{P}_j \in C_i\}$.*[59]

This does solve the instability problem just identified—but notice that it walks back much further from Reflection than is needed to do so. It replaces *every* point $\mathcal{P}_j$ that Reflection told $\mathcal{P}_i$ to be pulled towards with its informed version, $\widehat{\mathcal{P}}_j$. But the reason Reflection was unstable was not because each candidate $\mathcal{P}_i$ was pulled towards the others $\mathcal{P}_j$—it was that $\mathcal{P}_i$ *had nothing distinct from itself to pull back with*. Thus we didn't need to replace *all* the $\mathcal{P}_j$ with their informed selves; it suffices to replace $\mathcal{P}_i$ with *its* informed self $\widehat{\mathcal{P}}_i$ to make it so that $\mathcal{P}_i$ can stably give weight to the other candidates without being pulled into their convex hull. To see this, notice that in Figure 6, $\mathcal{P}_1$ *is* in the convex hull of $\widehat{\mathcal{P}}_1$ (left corner of the triangle) with the other points, as delineated by the red dotted line. This is just to say that $\mathcal{P}_i$ is in the convex hull of $\widehat{\mathcal{P}}_1$ and C_1^-—i.e. that it is stably both modest and modestly informed.

Thus the constraint that $\mathcal{P}_i$ be modestly informed combines the insights of both Reflection and New Reflection. From Reflection, it takes the idea that $\mathcal{P}_i$ should be pulled towards the uninformed opinions of the other candidates $\mathcal{P}_j \in C_i^-$ it leaves open. From New Reflection, it takes the idea that in order to do so stably, $\mathcal{P}_i$ must pull back with some anchor point other than itself—in particular, with its *informed* self $\widehat{\mathcal{P}}_i$—which, since it's not sensitive to higher-order doubts, can stably pull back.

Nevertheless, we may wonder why it has to be $\mathcal{P}_i$'s informed self ($\widehat{\mathcal{P}}_i$) in particular that serves as the anchor, as opposed to some other point more extreme than C_i^-. To see why this is so, consider one final way to reach the constraint that $\mathcal{P}_i$ be modestly informed. In particular, notice that Fact 4.2 entails that there's another formulation of the requirements of Reflection—namely, that $\mathcal{P}_i$ must be an average of its informed self and the (uninformed) opinions of the other candidates, where the weights in this average are extreme. That is, where $\lambda_{ij} \geq 0$ are non-negative weights that sum to 1, we have:

Corollary 4.4. *$\langle W, \mathcal{P}\rangle$ validates Reflection iff for each $i \in W$: $\mathcal{P}_i$ is modestly informed with extreme weights, i.e. $\mathcal{P}_i = \lambda_{ii}\widehat{\mathcal{P}}_i + \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathcal{P}_j$, with either $\lambda_{ii} = 1$ or $\lambda_{ii} = 0$.*[60]

Once we write Reflection this way, we can see that there are two completely discontinuous ways to satisfy it: the first is to have your opinions match your informed opinions (if $\lambda_{ii} = 1$); the second is to have them match an average of the other candidates' uninformed opinions (if $\lambda_{ii} = 0$). Once we see this bifurcation, it's natural to generalize it by allowing *intermediate* averages between these two extremes—giving some weight to your informed self and some weight to the other candidates uninformed opinions ($0 < \lambda_{ii} < 1$). That generalization is simply the requirement that $\mathcal{P}_i$ be modestly informed:

Corollary 4.5. *$\langle W, \mathcal{P}\rangle$ validates Total Trust iff for each $i \in W$, $\mathcal{P}_i$ is modestly informed: $\mathcal{P}_i = \lambda_{ii}\widehat{\mathcal{P}}_i + \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathcal{P}_j$ with $0 \leq \lambda_{ii} \leq 1$.*[61]

In short, Value retains the core idea of Reflection—namely, that your opinions should be pulled between your own informed opinions and the other candidates uninformed ones—but generalizes it to allow the balance between these two forces to be intermediate.

5 | OPEN QUESTIONS

Many of the things we'd like to defer to—people, evidence, chances—can be unsure whether they are worthy of deference. In such contexts the standard theories of deference break down. We've proposed Value as new theory of deference: you defer to an expert if you'd always prefer for them to make decisions on your behalf—in a slogan, *deferring opinions is deferring decisions*. Following Dorst (2020a), we observed that this theory is equivalent to the standard theories (Reflection and New Reflection) in the context of immodesty, but it both allows modesty (unlike Reflection) and rules out deference to anti-experts or Dutch-bookable ones (unlike New Reflection) (§1). However, we also showed that we lacked a general theory of modest Value (§2).

The point of this paper has been to give one. As we've seen, you value an expert iff you totally trust their estimates (§2), iff you expect their estimates to be more accurate than yours on any reasonable way of measuring accuracy (§3), iff all the expert candidates' opinions can be factored into their informed opinions along with their higher-order doubts, and your opinions are an average of theirs (§4). Collecting our theorems, we have the following characterization:

Theorem 5.1 (Characterization of Value)**.** *The following are equivalent:*

- *π values $\langle W, \mathcal{P}\rangle$;*
  For any $\mathcal{O}$, if S is recommended for $\mathcal{O}$, then $\forall O \in \mathcal{O}$, $\mathbb{E}_\pi(S) \geq \mathbb{E}_\pi(O)$;
  There's no fixed-option Dutch Book against transitioning from π to P.
- *π totally trusts $\langle W, \mathcal{P}\rangle$:*
  $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) \geq t$, for all X;
  $\pi(\cdot|P \in B) \in B$ for any biconvex B.
- *π epistemically values $\langle W, \mathcal{P}\rangle$:*
  $\mathbb{E}_\pi(I_X(P)) \leq \mathbb{E}_\pi(I_X(\pi))$, for all X and gsp I_X.
- *π is in the convex hull of C_π, and each $\mathcal{P}_i \in C_\pi$ is modestly informed:*

$$\pi = \sum_{\mathcal{P}_i \in C_\pi} \lambda_i \mathcal{P}_i \text{ and for each } \mathcal{P}_i \in C_\pi : \mathcal{P}_i = \lambda_{ii}\widehat{\mathcal{P}}_i + \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathcal{P}_j.$$

Theory in hand, we can draw out both formal lessons and philosophical questions.

The most obvious formal lesson is this. In the context of immodesty, deference can seem simple—all the plausible theories (Reflection, New Reflection, Value, Trust, etc.) coincide. But once we allow modest experts, important differences emerge. We've argued that probing these differences singles out Value as the most plausible, general account of deference. Nevertheless, within the space that modesty opens up, there's clearly a wide range of different deference principles that deserve to be explored (cf. Dorst 2020b). We hope that the methods we've found so useful here (asking computers; drawing pictures) will also be helpful in such further explorations.

Turning to philosophical questions, we'd like to focus on two. First, there are a variety of things that most philosophers agree we should defer to: objective chances (Lewis 1980; Levinstein 2019); our rational and more-informed future selves (van Fraassen 1984; Salow 2018); our own evidence (Good 1967; Dorst 2020a); and so on. If we are right, that means we must *value* such experts—

which in turn imposes structural constraints on what their opinions might be. For example, to value an expert, their opinions must obey *positive access*: if the expert is certain of q, they must be certain that the expert is certain of q; if $P(q) = 1$, then $P(P(q) = 1) = 1$ (Dorst 2020a, Fact 8.2). That raises a question: what substantive facts about these various experts explain why their opinions obey such structural features? What is it about objective chances (or rational credences, or evidential probabilities) that guarantees that they obey (say) positive access—and which theories of them can deliver this result? Thus theories of deference can be used to impose adequacy conditions on substantive accounts of chance (C. Dorst 2019; Gallow 2019a), diachronic rationality (cf. Schoenfield 2016b; Gallow 2019b), evidence (cf. Lasonen-Aarnio 2019; Das 2020a), and so on.

Second, every theory we've explored here is a theory of *global* deference—of what it takes to defer to an expert on *all* your opinions. As we mentioned at the beginning, this is an appropriate notion of deference for some experts (present chances, your current evidence) but not for others. We defer to Nate Silver about who'll win the election, but not about what the weather will be; we defer to our political opponents about what their favorite news network says, but not about whether it's correct; we defer to our future self about how busy we'll be next month, but not about how much we should work today; and so on. Most real-world deference is *local* deference: we defer to an expert's opinions about some questions but not others. Such limited deference is clearly both pervasive and philosophically important.

It's also formally interesting, for it turns out that question-sensitivity adds even more variability to our growing gamut of deference principles. Thinking of a question Q as a partition of logical space (Hamblin 1976; Roberts 2012), we can relativize all our principles to such questions. You reflect an expert *with respect to* Q iff for any partial answer to Q, you adopt the expert's credence in that answer upon learning what it is.[62] You totally trust an expert *with respect to* Q iff for any quantity whose values are determined by the answer to Q, you have a high estimate for that quantity upon learning that their estimate is high.[63] You value an expert *with respect to* Q iff for any decision-problem whose utilities are determined by the answer to Q, you'd like to give the expert power of attorney for that decision.[64]

Once we add such question-sensitivity, things get even more interesting. For example, you totally trust a frame *with respect to every 2-cell question* ($Q = \{q, \neg q\}$) iff you simply trust it (with respect to *every* question). As we've seen in Theorem 3.2, you totally trust an expert with respect to a question iff you expect their estimates of quantities determined by that question to be more accurate than your own on every reasonable way of measuring accuracy. We conjecture that, similarly, you totally trust an expert with respect to a question iff you value them with respect to that question.[65]

Finally—and perhaps most importantly—it turns out that you can totally trust, value, and epistemically value an *immodest* expert with respect to Q without reflecting them with respect to Q.[66] This means that even in the context of immodesty, the epistemic and pragmatic incentives to defer to someone do *not* always suffice to justify Reflection—they sometimes justify only Total Trust. That opens up further questions. How exactly do the local versions of Reflection and Total Trust relate? Is the latter indeed equivalent to (local) Value? And, more generally, how will the various theories of deference stack up once we turn our attention to the incredibly-common but under-explored domain of local, question-relative deference?

We don't know. So we should end on a modest note: although we think that our theory of deference does better than those currently on offer, there remains plenty of room for deference to be done better yet.

ACKNOWLEDGMENTS

This project benefited from lots of help along the way. Special thanks to Catrin Campbell-Moore, for sharing her characterization of inaccuracy measures with us, and to both her and Daniel Rothschild, for helping us to give an elementary proof of one of the core results (Theorem 3.2). Thanks to Nilanjan Das, for suggesting we look closer at the (it turns out, all-important) hyperplane separation theorem, and for giving detailed feedback on the paper once we had. Thanks to Jason Konek, Richard Pettigrew, Itai Sher, Isaac Wilhelm, and the Twitter math hive mind for help with a variety of places that we got stuck along the way. And thanks to Catrin Campbell-Moore (again!), Adam Elga, Dmitri Gallow, Harvey Lederman, Daniel Rothschild (again!), Ginger Schultheis, Jack Spencer, Timothy Williamson, and Snow Zhang for helpful feedback on earlier drafts. This work benefitted from the Mellon Foundation, New Directions fellowship 1905-06835.

ENDNOTES

[1] Letting 'P' be a definite description for the expert's opinions, whatever they are, the expert is modest iff they are less than certain that they are indeed the expert: there is some q, t such that $P(q) = t$ but $P(P(q) = t) < 1$. We follow Elga (2013) in the 'modest' terminology; note that it is orthogonal to the sense of 'immodesty' (or 'strict propriety') used as a constraint on accuracy scoring rules (Lewis 1971; Gibbard 2008; Joyce 2009; Pettigrew 2016; Horowitz 2018). See §3 for discussion.

[2] We here use 'Reflection' to refer to a variety of principles of the same structure; in this sense, the principles discussed by Miller (1966); Lewis (1980); Skyrms (1980); van Fraassen (1984); Gaifman (1988); Williamson (2000); Christensen (2007, 2010); Briggs (2009a, 2009b); Roush (2009, 2016); Elga (2013) and Mahtani (2017) are all Reflection principles. The tension with modesty is already noted by Lewis (1980); Elga (2013) gives a particularly clear and general explanation.

[3] We're following Pettigrew and Titelbaum (2014) in framing our problem as finding the right *form* of a deference principle across these various domains. We don't claim that there's a pretheoretic notion of 'deference' to analyze; rather, we use 'deference' as a name for the relation that should hold between your opinions and (at least some of) the various expert probabilities discussed above.

[4] See Hall (1994); Lewis (1994); Elga (2007, 2013); Pettigrew and Titelbaum (2014); Gallow (2019b); Christensen (2020). For different criticisms New Reflection, see Lasonen-Aarnio (2015).

[5] Available at https://www.kevindorst.com/DDB_notebook.html. The code was largely written by K.D. As you'll notice, he's no programmer. But it works.

[6] W is a finite set of worlds, and $\mathcal{P}$ is a function from worlds w to probability functions $\mathcal{P}_w$ that are defined over the subsets of W. **Convention**: technical terms are bolded when defined; the definitions are collected in Appendix A; bolded symbols always have the same meaning as their unbolded counterparts.

[7] This formalism assumes that at each world there is a unique, precise probability function that models the expert's opinions (cf. White 2005; Schoenfield 2012, 2014; Schultheis 2018; Carr 2019b).

[8] Precisely: $\mathcal{P}_w(P = \mathcal{P}_v) > 0$ for $\mathcal{P}_v \neq \mathcal{P}_w$; i.e. there's a $v \in W$ such that $\mathcal{P}_w(v) > 0$ and $\mathcal{P}_v \neq \mathcal{P}_w$.

[9] When the relevant opinions are themselves candidates for being the expert, we don't need this additional structure, since π will be guaranteed to be one of the $\mathcal{P}_w$ (Dorst 2020a, fn. 14). But in some cases the deferring opinions will not be candidate experts: you might know that, whoever the smartest and best-informed person is, it definitely isn't you.

[10] Here and throughout we leave implicit the restriction that the conditional probability is well-defined.

[11] Reflection has also been proposed as specifying the conditions under which shifting opinions from π to P counts as a "genuine learning experience" (Jeffrey 1988; Graves 1989; Skyrms 1990, 1997; Myrvold 2012; Huttegger 2013, 2014, 2017). The above suggests that this is too demanding, since one can surely regard the shift to some new opinions as a genuine learning experience even when those new opinions are modest (i.e., in this context, less than perfectly introspective). We think, instead, that the transition represents a genuine learning experience only if π *defers* to P—where deference, as we'll argue, need not require Reflection.

[12] Why are the two versions equivalent? Take any $w \in [\widehat{P} = \rho]$, so $\widehat{\mathcal{P}}_w := \mathcal{P}_w(\cdot|P = \mathcal{P}_w) = \rho$, and hence $\rho(P = \mathcal{P}_w) = 1$. Thus if $\mathcal{P}_x \neq \mathcal{P}_w$, then $\widehat{\mathcal{P}}_x := \mathcal{P}_x(\cdot|P = \mathcal{P}_x)$, so $\widehat{\mathcal{P}}_x(P = \mathcal{P}_w) = 0$, and $\widehat{\mathcal{P}}_x \neq \rho$. On the other hand since if $\mathcal{P}_x = \mathcal{P}_w$ then $\widehat{\mathcal{P}}_x = \rho$, we have that $[\widehat{P} = \rho] = [P = \mathcal{P}_w]$. Thus $\pi(\cdot|\widehat{P} = \rho) = \rho$ iff $\pi(\cdot|P = \mathcal{P}_w) = \rho = \widehat{\mathcal{P}}_w = \mathcal{P}_w(\cdot|P = \mathcal{P}_w)$; see Stalnaker (2019) and Dorst (2019b) for discussion.

[13] Precisely: $\mathcal{P}_a(a|P = \mathcal{P}_a) = 1 = \pi(a|P = \mathcal{P}_a)$ and $\mathcal{P}_b(b|P = \mathcal{P}_b) = 1 = \pi(b|P = \mathcal{P}_b)$.

[14] Here it's important to note that 'π' is a rigid designator for a fixed and known probability function, rather than a definite description for (e.g.) 'Your opinions, whatever they are', which, like 'P', could refer to different probability functions at different worlds.

[15] Value is related to a substantial literature arising out of Good (1967) about the conditions under which we should obtain additional information before deciding (e.g. Blackwell 1953; Savage 1954; Brown 1976; Skyrms 1990; Oddie 1997; Williamson 2000; Kadane et al. 2008; Myrvold 2012; Buchak 2013; Huttegger 2014; Bradley and Steele 2016; Ahmed and Salow 2018; Campbell-Moore and Salow 2019; Das 2020a; Pettigrew 2020). See Salow (2020) for a philosophical overview. Working in this tradition, Geanakoplos (1989) was effectively the first to investigate the possibility of satisfying Value when probabilities are modest, although he doesn't formulate his project in this way. He works within a sub-class of probability frames known as *prior frames* (endnote 25), thus our characterization (Theorem 4.1) will generalize his Theorem 1.

[16] By total expectation and then Reflection, we have $\mathbb{E}_\pi(S) = \sum_w \pi(P = \mathcal{P}_w)\mathbb{E}_\pi(S|P = \mathcal{P}_w) = \sum_w \pi(P = \mathcal{P}_w)\mathbb{E}_w(S)$. Since Reflection implies that the frame is immodest, $\mathbb{E}_w(S) = \mathbb{E}_w(S_w) \geq \mathbb{E}_w(O)$, meaning that the above sum is at least as great at $\sum_w \pi(P = \mathcal{P}_w)\mathbb{E}_w(O) = \sum_w \pi(P = \mathcal{P}_w)\mathbb{E}_\pi(O|P = \mathcal{P}_w) = \mathbb{E}_\pi(O)$. See Skyrms (1990); Huttegger (2014).

[17] Note that π values a frame only if it new-reflects it (see endnote 19). So if the frame is immodest, then $\mathcal{P}_i = \widehat{\mathcal{P}}_i$, so π new-reflects the frame iff it reflects it, and hence π values it only if it reflects it; endnote 16 proves the converse. Again see Skyrms (1990); Huttegger (2014).

[18] This frame also shows that Value does not entail another common weakening of Reflection—namely, that your credence must equal your best estimate of the expert's credence: $\pi(q) = \mathbb{E}_\pi(P(q))$ (Ismael 2008, 2015; Salow 2018, 2019; Gallow 2019b). For $\pi(a) = 0.5$, yet $\mathbb{E}_\pi(P(a)) = \pi(a) \cdot 0.9 + \pi(b) \cdot 0.2 = 0.55$. Unlike Reflection, such "estimate-matching" principles do not entail Value—note, for instance, that in our anti-expert frame in Figure 2, we *do* have $\pi(q) = \mathbb{E}_\pi(P(q))$ for all q.

[19] If New Reflection fails, we can make a conditional bet which the frame recommends taking but π doesn't want to take. Precisely: if (WLOG) $\pi(q|P = \mathcal{P}_w) < t < \mathcal{P}_w(q|P = \mathcal{P}_w)$, then let O_0 yield 0 everywhere and O_1 be a conditional bet which yields 0 if $P \neq \mathcal{P}_w$, $1 - t$ if $q \wedge [P = \mathcal{P}_w]$, and $-t$ if $\neg q \wedge [P = \mathcal{P}_w]$. Then $\mathcal{P}_w$ takes the bet, and every option has 0 utility when $P \neq \mathcal{P}_w$, so $\mathbb{E}_\pi(S) = \pi(P \neq \mathcal{P}_w)0 + \pi(P = \mathcal{P}_w)\mathbb{E}_\pi(O_1|P = \mathcal{P}_w)$, and $\mathbb{E}_\pi(O_1|P = \mathcal{P}_w) < t(1 - t) + (1 - t)(-t) = 0$, so $\mathbb{E}_\pi(S) < 0 = \mathbb{E}_\pi(O_0)$; Value fails.

[20] Precisely, it is a pair of decision problems $\mathcal{O}_1$ and $\mathcal{O}_2$ that both include a constant ("no bet") $O_0 = 0$ option, where $O \in \mathcal{O}_1$ maximizes expectation amongst $\mathcal{O}_1$ relative to π and S is a strategy recommended for $\mathcal{O}_2$ by P, such that $O(w) + S_w(w) < 0$ at every $w \in W$.

[21] Suppose π doesn't value P. Then there is a decision problem $\mathcal{O}$, an option $O \in \mathcal{O}$ and a strategy S recommended by P for $\mathcal{O}$ such that $\mathbb{E}_\pi(O) > \mathbb{E}_\pi(S)$. Let $\mathcal{O}_1 = \{O_0, O - S - \epsilon\}$ for $0 < \epsilon < \mathbb{E}_\pi(O) - \mathbb{E}_\pi(S)$, and let $\mathcal{O}_2 = \{O' - O : O' \in \mathcal{O}\}$. Note that $O_0 = O - O \in \mathcal{O}_2$, so both decision problems include a "no bet" option. Then $\mathbb{E}_\pi(O - S - \epsilon) = \mathbb{E}_\pi(O) - \mathbb{E}_\pi(S) - \epsilon > 0$, so $O - S - \epsilon$ maximizes expectation from $\mathcal{O}_1$ relative to π. Moreover, since $E_w(O' - O) = E_w(O') - E_w(O)$, and S is recommended for $\mathcal{O}$, $\mathbb{E}_w(S_w - O) \geq \mathbb{E}_w(O' - O)$ for every $O' \in \mathcal{O}$; so the strategy S' such that $S'_w = S_w - O$ is recommended for $\mathcal{O}_2$ by P. But for any world w: $O(w) - S(w) - \epsilon + S'_w(w) = O(w) - S_w(w) - \epsilon + S_w(w) - O(w) = -\epsilon < 0$; the combined course of action guarantees a loss.

[22] An expected loss implies that $\mathbb{E}_\pi(O + S) = \mathbb{E}_\pi(O) + \mathbb{E}_\pi(S) < 0$ for some O with maximal π-expectation. But by definition, $\mathbb{E}_\pi(O) \geq \mathbb{E}_\pi(O_0) = 0$, and by Value, $\mathbb{E}_\pi(S) \geq \mathbb{E}_\pi(O_0) = 0$.

Note: the formulation of these claims in terms of *fixed* options touches on a subtle but important point. Williamson (2000, Ch. 10) shows that many modest transitions—including some that Geanakoplos (1989) and Dorst (2020a) show to satisfy Value—are such that you can be offered a fixed set of *bets* such that, if at the later time you pay for each bet the maximal price that P is willing to pay, you should expect a loss. Das (2020a) generalizes this result, and strengthens it to show that you can also be forced into sure losses in this fashion. The way to reconcile these results with the above is that when we fix a bet but vary how much you pay for it across worlds, we are (in the technical sense) giving you different *options* $\mathcal{O}$ at different worlds: a bet that pays out \$1 if q and \$0 if $\neg q$, but which you must you pay \$0.60 for at worlds w where $\mathcal{P}_w(q) = 0.61$ and \$0.70 for at worlds x where $\mathcal{P}_x(q) = 0.71$ is not a single option, but is instead $O_1 = \begin{cases} 0.40 \text{ if} q \\ -0.60 \text{ if } \neg q \end{cases}$ at w and $O_2 = \begin{cases} 0.30 \text{ if } q \\ -0.70 \text{ if } \neg q \end{cases}$ at x.

The philosophical import of this fact is subtle. We are inclined to think that it casts doubt on the claim that such (non-fixed-option) Dutch books demonstrate irrationality, but this is an issue that deserves more discussion.

23 For example, note that in Figure 1, $[P(a) \geq 0.7] = \{a\}$ and $[P(b) \geq 0.6] = \{b\}$. Thus for any π: $\pi(a|P(a) \geq 0.7) = 1 \geq 0.7$, and $\pi(b|P(b) \geq 0.6) = 1 \geq 0.6$; learning which world the expert favors should lead you to favor that world as well—just even more so.

24 It's tempting to *combine* the two conditions to infer that Simple Trust implies that $P(q|[P(q) \geq t] \wedge [P(q) \leq t]) = t$, but Simple Trust does not imply this due to the non-monotonicity of probabilistic support (Dorst 2020a, §4.1).

25 The special subclass is the class of *prior frames*. A prior frame $\langle W, \mathcal{E}, \pi \rangle$ consist of a set of worlds W, a function $\mathcal{E}$ from worlds w to sets of worlds $\mathcal{E}_w$, and a probability distribution π that's regular over W. We can then recover a probability frame by defining $\mathcal{P}_w = \pi(\cdot \mid \mathcal{E}_w)$. Informally, we can think of $\mathcal{E}_w$ as the set of worlds which the expert's evidence at w fails to rule out, and of π as representing an initial probability distribution indicating what the expert takes the various bodies of evidence to support. On such an interpretation, prior frames allow for uncertainty about what evidence the expert has, but not for uncertainty about what they take these bodies of evidence to support. Prior frames are the focus in much of the related work (e.g. Geanakoplos 1989; Williamson 2000 2014, 2018; Cresto 2012; Lederman 2015; Lasonen-Aarnio 2015; Campbell-Moore 2016; Salow 2018, 2019; Das 2020a, 2020b), in part because they are more tractable. Nevertheless, there are compelling reasons to consider the wider class of probability frames. Prior frames build in a wide range of modeling assumptions about the candidate experts—for example, that they all share the same prior and all update by conditioning that prior on propositional evidence. On some interpretations (e.g. P = 'my present evidence') these assumptions are highly controversial; on others (e.g. P = 'the opinions of the smartest person in the room'), they are simply wrong. So when we can do without these assumptions, we should. Nevertheless, prior frames remain a useful, tractable starting point for most investigations.

26 The `checkTrust` function asks whether for each $\mathcal{P}_i$ in a given frame, $\mathcal{P}_i$ trusts the frame. To check that π trusts the frame (as well as that each $\mathcal{P}_i$ in the frame trusts the frame), one should enter the whole structure as if π is a world in the frame that always gets 0 probability, like so: $\begin{pmatrix} 0 & 0.17 & 0.56 & 0.27 \\ 0 & 0.45 & 0.10 & 0.45 \\ 0 & 0.15 & 0.70 & 0.15 \\ 0 & 0.30 & 0.10 & 0.60 \end{pmatrix}$.

27 $\mathbb{E}_\pi(\boldsymbol{X}) := \sum_w \pi(w)X(w)$. Note that (total expectation:) for any partition Q, $\mathbb{E}_\pi(X) = \sum_{q \in Q} \pi(q)\mathbb{E}_\pi(X|q)$, where $\mathbb{E}_\pi(X|q) = \sum_w \pi(w|q)X(w)$.

28 For aficionados: this principle can be re-stated in terms of convex sets; see Total Trust (convexity version) below.

29 That is, Total Trust implies that for any X, Y: $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq \mathbb{E}(Y)) \geq \mathbb{E}_\pi(Y|\mathbb{E}(X) \geq \mathbb{E}(Y))$. Note that by linearity of expectations, $\mathbb{E}(X) \geq E(Y)$ iff $\mathbb{E}(X - Y) \geq 0$; hence $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq \mathbb{E}(Y)) < \mathbb{E}_\pi(Y|\mathbb{E}(X) \geq \mathbb{E}(Y))$ implies $\mathbb{E}_\pi(X - Y|\mathbb{E}(X - Y) \geq 0) < 0$, violating Total Trust.

30 For more difficult theorems we include only proof sketches in the main text; full proofs can be found in Appendix B.

31 Indeed, if there's one methodological moral that we take away from this paper, it's that we should've long ago taken to heart the advice of Hanti Lin and Kevin Kelly (Lin and Kelly 2012a, 2012b) and *drawn pictures* of probability frames. This was what led us to discover most of the results to come. In fact, arguably one of the reasons that characterizations within prior frames (endnote 25) have come so much more easily than in probability frames is that the former are easier to draw.

32 If, looking at this diagram, you are thinking "why not just say that π is in the convex hull of (within the triangle circumscribed by) $\{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\}$?", the answer is that this condition is necessary but not sufficient for Value—as proven below in Theorem 4.1 and illustrated by Figure 2 above.

33 Formally: C is convex if $\delta \in C$ whenever $\delta = \sum_i \lambda_i \rho_i$ for some $\rho_1, \ldots, \rho_n \in C$ with $\lambda_i \geq 0$ such that $\sum_i \lambda_i = 1$.

34 Precisely: if both B and B^c are convex, then by the hyperplane separation theorem, their boundary is a hyperplane. Recall that we are restricting attention to finite probability frames, so we don't need to worry about strengthening the $\mathbb{E}(X) \geq t$ condition in Total Trust to deal with edge cases. In particular, $C = \{\mathcal{P}_w : \mathcal{P}_w \in B\}$ and $D = \{\mathcal{P}_w : \mathcal{P}_w \notin B\}$ are both finite, so their convex hulls are closed and compact, meaning that there is a hyperplane *strongly* separating them: $\exists Y, t, \varepsilon > 0$ such that $C \subset \{\rho : \mathbb{E}_\rho(Y) \geq t\}$ and $D \subset \{\rho : \mathbb{E}_\rho(Y) \leq t - \varepsilon\}$.

35 There may be more to be said about why the idea behind Total Trust does not motivate any stronger principles. One hypothesis builds on the "directed" nature of the set of probability functions above a cut. Intuitively, $\{\rho : \rho(q) \geq t\}$ is directed towards q-worlds and away from $\neg q$ worlds. Indeed, this can be made precise and

generalized: any set $A = \{\rho : \mathbb{E}_\rho(X) \geq t\}$ has a unique direction associated with it—namely, that of the vector X. Moreover, any (non-empty) such set will contain at least one extreme point (world) of the simplex, and the angle between X and any extreme point inside A is less than that between X and any extreme point outside of A. Thus any such A is directed at certain worlds and away from others. The same is not true of sets that are not determined by one of our cuts, even if they are convex. This might give a sense for why conditioning on such an A should effect a shift in a predictable direction, while conditioning on another set, even a convex one, is unconstrained—and hence why Total Trust should be true, but any generalization of it would overstep the mark the way Reflection does.

36 For any X, t and $q \subseteq W$: $\mathbb{E}_\pi(X|q \wedge [\mathbb{E}(X|q) \geq t]) \geq t$. Note that $B = \{\rho : \mathbb{E}_\rho(X|q) \geq t\}$ is biconvex. To see this, first note that if $\delta = \sum_i \lambda_i \rho_i$, then $\delta(\cdot|q)$ is also a convex mixture of the $\rho_i(\cdot|q)$: for any p, $\delta(p|q) = \frac{\delta(p \wedge q)}{\delta(q)} = \frac{\sum_i \lambda_i \rho_i(q) \frac{\rho_i(p \wedge q)}{\rho_i(q)}}{\sum_k \lambda_k \rho_k(q)} = \sum_i (\frac{\lambda_i \rho_i(q)}{\sum_k \lambda_k \rho_k(q)} \rho_i(p|q))$; so $\delta(\cdot|q) = \sum_i (\frac{\lambda_i \rho_i(q)}{\sum_k \lambda_k \rho_k(q)} \rho_i(\cdot|q))$. Note that this mixture gives positive weight only to the ρ_i for which $\rho_i(\cdot|q)$ is well-defined. Now, for any $\rho_i \in B$ and non-negative λ_i that sum to 1, $\sum_i \lambda_i \mathbb{E}_{\rho_i}(X|q) \geq \sum_i \lambda_i t = t$, so $\sum_i \lambda_i \rho_i \in B$. Likewise, for any $\delta_i \in B^c$, if $\mathbb{E}_{\delta_i}(X|q)$ is well-defined then it's less than t, so $\sum_i \lambda_i \mathbb{E}_{\delta_i}(X|q) < \sum_i \lambda_i t = t$, so $\sum_i \lambda_i \delta_i \in B^c$. Thus B is biconvex, so by Total Trust, we have that $\pi^* := \pi(\cdot|P \in B) \in B$, and thus $\mathbb{E}_\pi(X|q \wedge [\mathbb{E}(X|q) \geq t]) = \mathbb{E}_{\pi^*}(X|q) \geq t$.

37 For arbitrary q, suppose $\mathcal{P}_j(q|P = \mathcal{P}_j) = s$. Then $[P = \mathcal{P}_j] \wedge [P(q|P = \mathcal{P}_j) \geq s] = [P = \mathcal{P}_j]$ and similarly $[P = \mathcal{P}_j] \wedge [P(q|P = \mathcal{P}_j) \leq s] = [P = \mathcal{P}_j]$; so by Trust, $\pi(q|P = \mathcal{P}_j) = \pi(q|[P = \mathcal{P}_j] \wedge [P(q|P = \mathcal{P}_j) \geq s) \geq s$, and also $\pi(q|P = \mathcal{P}_j) = \pi(q|[P = \mathcal{P}_j] \wedge [P(q|P = \mathcal{P}_j) \leq s) \leq s$, so $\pi(q|P = \mathcal{P}_j) = s$. Since q was arbitrary, that means Trust implies New Reflection.

38 That is, $\sum_i \frac{1}{n} \pi(q_i | \sum_i \frac{1}{n} P(q_i) \geq t) \geq t$. Note that $B = \{\rho : \sum_i \frac{\rho(q_i)}{n} \geq t\}$ is a biconvex set, since if $\rho_j \in B$, then $\sum_i \frac{1}{n} \sum_j \lambda_j \rho_j(q_i) = \sum_j \sum_i \frac{1}{n} \lambda_j \rho_j(q_i) = \sum_j \lambda_j \sum_i \frac{\rho_j(q_i)}{n} \geq \sum_j \lambda_j t \geq t$, so $\sum_j \lambda_j \rho_j \in B$; by parallel reasoning B^c is also convex. Hence $\pi(\cdot|P \in B) \in B$, and the result holds.

39 E.g. Greaves and Wallace (2006); Joyce (2009); Predd et al. (2009); Pettigrew (2016). Note that this property is sometimes called "immodesty" (Lewis 1971)—terminology that is orthogonal to our own (endnote 1).

40 Precisely: let A be any strictly proper accuracy measure (so $\forall \rho, \delta : \mathbb{E}_\rho(A(\rho)) > \mathbb{E}_\rho(A(\delta))$ if $\delta \neq \rho$) and let the set of options be those which yield utility matching the accuracy of either π or some function in the frame $\mathcal{P}_i$: $\mathcal{O} = \{A(\rho) : \rho \in \{\pi\} \cup \{\mathcal{P}_w : w \in W\}\}$. (Recall that our frames our finite, so that this is a finite set of options.) Then by strict propriety, for any $i \in W$ and $\rho \neq \mathcal{P}_i$ in $\{\pi\} \cup \{\mathcal{P}_w : w \in W\}$, we have $\mathbb{E}_i(A(\mathcal{P}_i)) > \mathbb{E}_i(A(\rho))$, i.e. $\forall O \neq A(\mathcal{P}_i)$, $\mathbb{E}_i(A(\mathcal{P}_i)) > \mathbb{E}_i(O)$; hence the uniquely recommended strategy S is such that for all i, $S_i = A(\mathcal{P}_i)$, and S picks out "the accuracy of the expert credence function, whatever it is": for all w, $S(w) = A(\mathcal{P}_w, w)$. By Value, we know that $\mathbb{E}_\pi(A(\pi)) \leq \mathbb{E}_\pi(S) = \mathbb{E}_\pi(A(P))$. Since A was an arbitrary strictly proper measure, it follows that Value implies that π expects the expert to be at least as accurate as itself under any such measure.

41 See Rosenkrantz (1981); Schervish (1989); Oddie (1997); Joyce (1998, 2009); Greaves and Wallace (2006); Predd et al. (2009); Schoenfield (2016b, 2016a, 2017); Pettigrew (2016); Carr (2017, 2019a); De Bona and Staffel (2017); Levinstein (2017b); Campbell-Moore and Salow (2019, 2020); Campbell-Moore (2020); Campbell-Moore and Levinstein (2020); Konek and Levinstein (2019), and many others.

42 Precisely: if $|\delta(q) - i| < |\rho(q) - i|$, then $I_q(\delta, i) < I_q(\rho, i)$ for $i = 0, 1$.

43 Precisely: for any δ, ρ, $\mathbb{E}_\delta(I_q(\delta)) \leq \mathbb{E}_\delta(I_q(\rho))$ with equality iff $\delta(q) = \rho(q)$.

44 Precisely: there are local inaccuracy measures I_q such that for all w, δ, $I(\delta, w) = \sum_{q \subseteq W} I_q(\delta, \mathbb{1}_q(w))$.

45 See Levinstein (2019), Appendix A for the details.

46 Precisely: using the Schervish-style characterization from §B.3.2, letting $e = \mathbb{E}_\pi(X)$, $f(t) = \begin{cases} 1 \text{ if } t \in [-0.1, 0.1] \\ 0.001 \text{ otherwise} \end{cases}$ and $I_X(\pi, w) = \int_{\min(e, X(w))}^{\max(e, X(w))} |t - X(w)| f(\mathrm{d}t)$, then this is a generally strictly proper scoring rule and we have $\mathbb{E}_\pi(I_{O_1}(P)) = 1.107 > 1.082 = \mathbb{E}_\pi(I_{O_1}(\pi))$.

47 Precisely: for any probabilistic π, $\mathbb{E}_\pi(I_X(\pi)) < \mathbb{E}_\pi(I_X(s))$ whenever $\mathbb{E}_\pi(X) \neq s$.

48 Precisely: for any gsp I_X, $\mathbb{E}_\pi(I_X(P)) \leq \mathbb{E}_\pi(I_X(\pi))$, with equality only if $\pi(\mathbb{E}(X) = \mathbb{E}_\pi(X)) = 1$.

49 Appendix B.3 gives two proofs of this result, one using Campbell-Moore's (2020) Schervish-style characterization, and the other using purely elementary methods; the latter is heavily indebted to help from Catrin Campbell-Moore and Daniel Rothschild.

50 Let our frame be $\begin{pmatrix} 0.9 & 0.1 & 0 \\ 0.9 & 0.1 & 0 \\ 0.4 & 0.1 & 0.5 \end{pmatrix}$ and let $\pi = \mathcal{P}_3 = \begin{pmatrix} 0.4 & 0.1 & 0.5 \end{pmatrix}$. $\pi(w_1|P(w_s) \geq 0.9) = 0.8$ and $\pi(w_1|P = \mathcal{P}_1) = 0.8 \neq 0.9 = \mathcal{P}_1(w_1|P = \mathcal{P}_1)$, so π neither (simply) trusts nor new-reflects this frame. Nevertheless, for every $q \subseteq W$, π expects P to have a better Brier score with respect to q than itself.

51 Recall that Figure 2 shows that π can new-reflect P while knowing that P is less accurate than π. And letting $\pi = \begin{pmatrix} 0.8 & 0.1 & 0.1 \end{pmatrix}$ and $\langle W, \mathcal{P} \rangle = \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ yields an example in which π expects P to be more accurate than it on every random variable, according to squared Euclidean distance, yet $\pi(P(\{w_2, w_3\}) \geq 0.4) = 1$ while $\pi(\{w_2, w_3\}) = 0.2 < 0.4$, so Simple Trust fails. Thanks to Richard Pettigrew for the example.

52 Precisely, for $X = \begin{pmatrix} 5 & -1 & -10 \end{pmatrix}$, the purple region is $\{\rho : \mathbb{E}_\rho(X) \geq 0\}$, so $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq 0) < 0$.

53 Precisely, $\mathbf{CH}(\{\rho_1, \ldots, \rho_n\}) = \{\delta : \exists \lambda_i \geq 0 \text{ and } \sum \lambda_i = 1 \text{ such that } \delta = \sum \lambda_i \rho_i\}$.

54 Precisely, let $\boldsymbol{C_\rho} := \{\delta : \rho(P = \delta) > 0\}$ be the set of Candidates that ρ thinks might be the expert (abbreviating $C_{\mathcal{P}_i}$ to $\boldsymbol{C_i}$ for $\mathcal{P}_i$ in the frame), and let $\boldsymbol{C_\rho^-} = C_\rho \setminus \{\rho\}$ be those candidates other than ρ itself. Then $\mathcal{P}_i$ is modestly informed iff $\mathcal{P}_i$ is in the convex hull of $\{\widehat{\mathcal{P}}_i\} \cup C_i^-$, i.e. iff there are non-negative weights λ_{ij} that sum to 1 such that $\mathcal{P}_i = \lambda_{ii}\widehat{\mathcal{P}}_i + \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathcal{P}_j$.

55 Or, if you prefer, as what the "first-order evidence" warrants at that world (Dorst 2019b).

56 Notably, while π reflects/values/totally-trusts $\langle W, \mathcal{P} \rangle$ only if every $\mathcal{P}_i \in W_\pi$ reflects/values/totally-trusts P, π can new-reflect $\langle W, \mathcal{P} \rangle$ even when $\mathcal{P}_i \in W_\pi$ doesn't new-reflect $\langle W, \mathcal{P} \rangle$. Thus the requirement that New Reflection is valid on a frame is stronger than the requirement that some π that assigns positive probability to all $w \in W$ new-reflects it. If one should defer only to those who defer to themselves, this is simply another argument that New Reflection is too weak for deference.

57 *Proof.*($\Rightarrow$:) If $\mathcal{P}_i(P = \mathcal{P}_i) > 0$ then $\mathcal{P}_i$ must be immodest since $\mathcal{P}_i(P = \mathcal{P}_i|P = \mathcal{P}_i) = 1$. If not, then $C_i = C_i^-$, so by total probability and Reflection $\mathcal{P}_i = \sum_{\mathcal{P}_j \in C_i^-} \mathcal{P}_i(P = \mathcal{P}_j)\mathcal{P}_j$.

($\Leftarrow$:) Let $A := \{\mathcal{P}_i : i \in W\}$, let B be the extreme points of A's convex hull. Every $\mathcal{P}_i \in B$ must be immodest, since if $\mathcal{P}_i$ is an extreme point then it's not in $CH(A - \{\mathcal{P}_i\})$, and so not in the convex hull of C_i^-; so Reflection holds throughout B. For $\mathcal{P}_j \in A - B$, $\mathcal{P}_j$ is in the convex hull of the immodest B, so $\mathcal{P}_j(P \in B) = 1$, hence $\mathcal{P}_j = \sum_{\mathcal{P}_i \in B} \lambda_i \mathcal{P}_i$. Taking any q and $\mathcal{P}_k$, we have $\mathcal{P}_j(q|P = \mathcal{P}_k) = \frac{\mathcal{P}_j(q \wedge [P = \mathcal{P}_k])}{\mathcal{P}_j(P = \mathcal{P}_k)} = \frac{\sum_i \lambda_i \mathcal{P}_i(q \wedge [P = \mathcal{P}_k])}{\sum_i \lambda_i (P = \mathcal{P}_k)} = \frac{\lambda_k \mathcal{P}_k(q \wedge [P = \mathcal{P}_k])}{\lambda_k \mathcal{P}_k(P = \mathcal{P}_k)} = \mathcal{P}_k(q)$. (The last two equalities comes from the fact that each $\mathcal{P}_i \in B$ is immodest.) Hence Reflection holds throughout $A - B$.

58 By total probability and then Reflection, $\mathcal{P}_i = \sum_{\mathcal{P}_j \in C_i} \mathcal{P}_i(P = \mathcal{P}_j)\mathcal{P}_i(\cdot|P = \mathcal{P}_j) = \sum_{\mathcal{P}_j \in C_i} \mathcal{P}_i(P = \mathcal{P}_j)\mathcal{P}_j = \mathcal{P}_i(P = \mathcal{P}_i)\mathcal{P}_i + \sum_{\mathcal{P}_j \in C_i^-} \mathcal{P}_i(P = \mathcal{P}_j)\mathcal{P}_j$.

59 *Proof.*($\Rightarrow$:) If $\mathcal{P}_i$ satisfies New Reflection, then by total probability and then New Reflection we have $\mathcal{P}_i = \sum_{\mathcal{P}_j \in C_i} \mathcal{P}_i(P = \mathcal{P}_j)\mathcal{P}_i(\cdot|P = \mathcal{P}_j) = \sum_{\mathcal{P}_j \in C_i} \mathcal{P}_i(P = \mathcal{P}_j)\widehat{\mathcal{P}}_j$, so $\mathcal{P}_i$ is in the CH of $\{\widehat{\mathcal{P}}_j; \mathcal{P}_i \in C_i\}$.

($\Leftarrow$:) Suppose $\mathcal{P}_i = \sum_{\mathcal{P}_j \in C_i} \lambda_j \widehat{\mathcal{P}}_j$. Taking any q and $\widehat{\mathcal{P}}_k$ for which $\mathcal{P}_i(\cdot|\widehat{P} = \widehat{\mathcal{P}}_k)$ is defined, we have $\mathcal{P}_i(q|\widehat{P} = \widehat{\mathcal{P}}_k) = \frac{\mathcal{P}_i(q \wedge [\widehat{P} = \widehat{\mathcal{P}}_k])}{\mathcal{P}_i(\widehat{P} = \widehat{\mathcal{P}}_k)} = \frac{\sum_j \lambda_j \widehat{\mathcal{P}}_j(q \wedge [\widehat{P} = \widehat{\mathcal{P}}_k])}{\sum_j \lambda_j (\widehat{P} = \widehat{\mathcal{P}}_k)} = \frac{\lambda_k \widehat{\mathcal{P}}_k(q \wedge [\widehat{P} = \widehat{\mathcal{P}}_k])}{\lambda_k \widehat{\mathcal{P}}_k(\widehat{P} = \widehat{\mathcal{P}}_k)} = \widehat{\mathcal{P}}_k(q)$. (The last two equalities come from the fact that each $\widehat{\mathcal{P}}_j$ is immodest.) Hence New Reflection (informed version) holds.

60 By Fact 4.2, each $\mathcal{P}_i$ is either immodest or in the convex hull of C_i^-. If the former, then $\mathcal{P}_i = \widehat{\mathcal{P}}_i$, so $\lambda_{ii} = 1$; and if the latter, then $\mathcal{P}_i = \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathcal{P}_j$, so $\lambda_{ii} = 0$.

61 By Theorem 4.1, $\mathcal{P}_i$ totally trusts the frame iff it's in the convex hull of C_i and C_i is modestly informed. If $\mathcal{P}_i$ assigns positive probability to itself, then it's automatically in the convex hull of $\mathcal{P}_i \in C_i$, and so the only requirement is that all of C_i (including $\mathcal{P}_i$) be modestly informed. If $\mathcal{P}_i$ does *not* assign positive probability to itself, then since $C_i = C_i^-$, it's in the convex hull of C_i iff it's modestly informed. Note: Whereas the weights λ_{ij} in Facts 4.2, 4.3, and 4.4 turn out to always equal $\mathcal{P}_i(P = \mathcal{P}_j)$, this will not in general be true for the weights λ_{ij} used in Corollary 4.5.

62 Precisely: if $q = \bigcup q_i$, for $q_i \in Q$, then $\pi(q|P = \rho) = \rho(q)$.

63 Precisely: if X is such that for all w, w' in the same Q-cell, $X(w) = X(w')$, then $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) \geq t$.

64 Precisely: if for all $O \in \mathcal{O}$, $O(w) = O(w')$ whenever w and w' are in the same Q-cell, then if S is recommended for $\mathcal{O}$, $\mathbb{E}_\pi(S) \geq \mathbb{E}_\pi(O)$, for any $O \in \mathcal{O}$.

[65] The right-to-left direction is true and easy to prove using the same reasoning as that in Lemma B.1; it's the left-to-right direction which is open. Levinstein (2017b) shows how scoring rules for individual propositions can be thought of as measuring the expected practical (dis)utility of having a certain credence in a given proposition when it is uncertain what decision problem you face whose outcome is determined solely by the truth-value of the proposition in question. Since Levinstein's (2019) result shows that Simple Trust is enough to expect an expert to be at least as accurate as you are for a given proposition (and therefore to have a credence with higher-expected utility), we believe it's likely the left-to-right direction holds as well. However, it is unclear how or whether the identification of propositional scoring rules with pragmatic expected disutility generalizes to gsp scoring rules.

[66] Let $\langle W, \mathcal{P}\rangle = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.6 & 0.4 & 0 \\ 0 & 0.6 & 0.4 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ and $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, let $q = \{w_1, w_2\}$ and $Q = \{q, \neg q\}$. Then $\pi(q|P(q) = 0.6) = \pi(q|\{2, 3\}) = 0.5$, so π doesn't reflect $\langle W, \mathcal{P}\rangle$ with respect to Q; and since the frame is immodest, π likewise does not new-reflect it with respect to Q. Nevertheless, π does totally trust $\langle W, \mathcal{P}\rangle$ with respect to Q, since, for instance, $\pi(q|P(q) \geq 0.6) = \pi(q|\{1, 2, 3\}) = {}^{2}/_{3}$ and $\pi(q|P(q) \leq 0.6) = \pi(q|\{2, 3, 4\}) = {}^{1}/_{3}$; and, in fact, π values this frame with respect to Q.

[67] If you're wondering why we said the Schervish-style proof is more illuminating, this scoring rule is one reason. We constructed this rule using the Campbell-Moore, Schervish-style characterization, setting $\lambda(\mathrm{d}t) = 2\mathrm{d}t$ everywhere except $[\alpha, \beta]$, where instead it is $2C\mathrm{d}t$. Then we made B.A.L. compute the integrals.

REFERENCES

Ahmed, Arif and Salow, Bernhard, 2018. 'Don't Look Now'. *British Journal for the Philosophy of Science*, To appear.

Blackwell, David, 1953. 'Equivalent Comparisons of Experiments'. *The Annals of Mathematical Statistics*, 24(2):265–272.

Bradley, Seamus and Steele, Katie, 2016. 'Can free evidence be bad? Value of information for the imprecise probabilist'. *Philosophy of Science*, 83(1):1–28.

Briggs, R., 2009a. 'Distorted Reflection'. *Philosophical Review*, 118(1):59–85.

Briggs, Ray, 2009b. 'The Anatomy of the Big Bad Bug'. *Nous*, 43(3):428–449.

Brown, Peter M, 1976. 'Conditionalization and expected utility'. *Philosophy of Science*, 43(3):415–419.

Buchak, Lara, 2013. *Risk and rationality*. Oxford University Press.

Campbell-Moore, Catrin, 2016. *Self-Referential Probability*. Ph.D. thesis.

_____, 2020. 'Accuracy, Estimates, and Representation Results'.

Campbell-Moore, Catrin and Levinstein, Benjamin A, 2020. 'Strict Propriety is Weak'. *Analysis*, To Appear.

Campbell-Moore, Catrin and Salow, Bernhard, 2019. 'Avoiding Risk and Avoiding Evidence'.

_____, 2020. 'Accurate Updating for the Risk-Sensitive'. *The British Journal for the Philosophy of Science*.

Carr, Jennifer Rose, 2017. 'Epistemic Utility Theory and the Aim of Belief'. *Philosophy and Phenomenological Research*, 95(3):511–534.

_____, 2019a. 'A modesty proposal'. *Synthese*, 1–21.

_____, 2019b. 'Imprecise Evidence without Imprecise Credences'. *Philosophical Studies*, To appear.

Christensen, David, 2007. 'Epistemic Self-Respect'. *Proceedings of the Aristotelian Society*, CVII(3):319–337.

_____, 2010. 'Rational Reflection'. *Philosophical Perspectives*, 24:121–140.

_____, 2020. 'Akratic (epistemic) modesty'. *Philosophical Studies*, To appear(September).

Cresto, Eleonora, 2012. 'A Defense of Temperate Epistemic Transparency'. *Journal of Philosophical Logic*, 41(6):923–955.

Das, Nilanjan, 2020a. 'Externalism and Exploitability'. *Philosophy and Phenomenological Research*, To Appear.

_____, 2020b. 'The Value of Biased Information'. *The British Journal for the Philosophy of Science*, To Appear.

De Bona, Glauber and Staffel, Julia, 2017. 'Graded Incoherence for Accuracy-Firsters'. *Philosophy of Science*, 84(2):189–213.

Dorst, Chris, 2019a. 'Towards a best predictive system account of laws of nature'. *The British Journal for the Philosophy of Science*, 70(3):877–900.

Dorst, Kevin, 2019b. 'Higher-Order Uncertainty'. In Mattias Skipper Rasmussen and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 35–61. Oxford University Press.

_____, 2020a. 'Evidence: A Guide for the Uncertain'. *Philosophy and Phenomenological Research*, 100(3):586–632.

_____, 2020b. 'Higher-Order Evidence'. In Maria Lasonen-Aarnio and Clayton Littlejohn, eds., *The Routledge Handbook for the Philosophy of Evidence*. Routledge.

Elga, Adam, 2007. 'Reflection and Disagreement'. *Noûs*, 41(3):478–502.

_____, 2013. 'The puzzle of the unmarked clock and the new rational reflection principle'. *Philosophical Studies*, 164(1):127–139.

Gaifman, Haim, 1988. 'A Theory of Higher Order Probabilities'. In Brian Skyrms and William L Harper, eds., *Causation, Chance, and Credence*, volume 1, 191–219. Kluwer.

Gallow, J. Dmitri, 2017. 'Local & Global Experts'.

_____, 2019a. 'A Subjectivist's Guide to Deterministic Chance'. *Synthese*, To Appear:1–34.

_____, 2019b. 'Updating for externalists'. *Noûs*, (November 2018):1–30.

Geanakoplos, John, 1989. 'Game Theory Without Partitions, and Applications to Speculation and Consensus'. Cowles Fou.

Gibbard, Alan, 2008. 'Rational Credence and the Value of Truth'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 2. Oxford University Press.

Good, I J, 1967. 'On the Principle of Total Evidence'. *The British Journal for the Philosophy of Science*, 17(4):319–321.

Graves, Paul R, 1989. 'The total evidence theorem for probability kinematics'. *Philosophy of Science*, 56(2):317–324.

Greaves, Hilary and Wallace, David, 2006. 'Justifying conditionalisation: conditionalisation maximizes expected epistemic utility Introduction: Justifying conditionalisation'. 1–23.

Hall, Ned, 1994. 'Correcting the Guide to Objective Chance'. *Mind*, 103(412):505–517.

Hamblin, Charles L, 1976. 'Questions in montague english'. In *Montague grammar*, 247–259. Elsevier.

Hintikka, Jaako, 1962. *Knowledge and Belief*. Cornell University Press.

Horowitz, Sophie, 2018. 'Epistemic value and the Jamesian goals'. *Epistemic Consequentialism*, 269–289.

Huttegger, Simon M, 2013. 'In defense of reflection'. *Philosophy of Science*, 80(3):413–433.

_____, 2014. 'Learning experiences and the value of knowledge'. *Philosophical Studies*, 171(2):279–288.

_____, 2017. *The probabilistic foundations of rational learning*. Cambridge University Press.

Ismael, J. T., 2015. 'In defense of IP: A response to pettigrew'. *Nous*, 49(1):197–200.

Ismael, Jenann, 2008. 'Raid! Dissolving the big, bad bug'. *Nous*, 42(2):292–307.

Jeffrey, Richard, 1988. 'Conditioning, kinematics, and exchangeability'. In *Causation, chance and credence*, 221–255. Springer.

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

_____, 2009. 'Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief'. In Franz Huber and Christoph Schmidt-Petri, eds., *Degrees of Belief*, 263–297. Springer.

Kadane, Joseph B, Schervish, Mark, and Seidenfeld, Teddy, 2008. 'Is ignorance bliss?' *The Journal of Philosophy*, 105(1):5–36.

Konek, Jason and Levinstein, Benjamin A., 2019. 'The Foundations of Epistemic Decision Theory'. *Mind*, 128(509):69–107.

Kripke, Saul A, 1963. 'Semantical analysis of modal logic i normal modal propositional calculi'. *Mathematical Logic Quarterly*, 9(5-6):67–96.

Lasonen-Aarnio, Maria, 2013. 'Disagreement and evidential attenuation'. *Nous*, 47(4):767–794.

_____, 2015. 'New Rational Reflection and Internalism about Rationality'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 5, 145–171. Oxford University Press.

_____, 2019. 'Higher-Order Defeat and Evincibility'. *Higher-Order Evidence: New Essays*, 144–171.

Lederman, Harey, 2015. 'People with Common Priors Can Agree to Disagree'. *The Review of Symbolic Logic*, 8(1):11–45.

Levinstein, B. A., 2019. 'Accuracy, Deference, and Chance'.

Levinstein, Ben, 2017a. 'Permissive Rationality and Sensitivity'. *Philosophy and Phenomenological Research*, XCIV(2):343–370.

Levinstein, Benjamin Anders, 2017b. 'A pragmatist's guide to epistemic utility'. *Philosophy of Science*, 84(4):613–638.

Lewis, David, 1971. 'Completeness and Decidability of Three Logics of Counterfactual Conditionals'. *Theoria*, 17:74–85.

_____, 1980. 'A subjectivist's guide to objective chance'. In Richard C Jeffrey, ed., *Studies in Inductive Logic and Probability*, volume 2. University of California Press.

_____, 1994. 'Humean Supervenience Debugged'. *Mind*, 103(412):473–490.

Lin, Hanti and Kelly, Kevin T, 2012a. 'A geo-logical solution to the lottery paradox, with applications to conditional logic'. *Synthese*, 186(2):531–575.

_____, 2012b. 'Propositional reasoning that tracks probabilistic reasoning'. *Journal of philosophical logic*, 41(6):957–981.

Mahtani, Anna, 2017. 'Deference, respect and intensionality'. *Philosophical Studies*, 174(1):163–183.

Miller, David, 1966. 'A paradox of information'. *The British Journal for the Philosophy of Science*, 17(1):59–61.

Myrvold, Wayne C, 2012. 'Epistemic values and the value of learning'. *Synthese*, 187(2):547–568.

Nissan-Rozen, Ittay and Spectre, Levi, 2019. 'A pragmatic argument against equal weighting'. *Synthese*, 196(10):4211–4227.

Oddie, Graham, 1997. 'Conditionalization, Cogency, and Cognitive Value'. *The British Journal for the Philosophy of Science*, 48(4):533–541.

Pettigrew, Richard, 2015. 'Risk, rationality and expected utility theory'. *Canadian Journal of Philosophy*, 45(5-6):798–826.

_____, 2016. 'Measuring Accuracy: A New Account'. In *Accuracy and the Laws of Credence*.

_____, 2020. 'On the pragmatic and epistemic virtues of inference to the best explanation'.

Pettigrew, Richard and Titelbaum, Michael G, 2014. 'Deference Done Right'. *Philosopher's Imprint*, 14(35):1–19.

Predd, J, Seringer, R, Lieb, E H, Osherson, D, Poor, H V, and Kulkarni, S, 2009. 'Probabilistic coherence and proper scoring rules'. *IEEE Transactions on Information Theory*, 55(10):4786–4792.

Roberts, Craige, 2012. 'Information structure in discourse: Towards an integrated formal theory of pragmatics'. *Semantics and Pragmatics*, 5(6):1–69.

Rosenkrantz, Roger D, 1981. *Foundations and Applications of Inductive Probability*. Ridgeview Press.

Roush, Sherrilyn, 2009. 'Second Guessing: A Self-Help Manual'. *Episteme*, 251–268.

_____, 2016. 'Knowledge of Our Own Beliefs'. *Philosophy and Phenomenological Research*, 93(3).

Salow, Bernhard, 2018. 'The Externalist's Guide to Fishing for Compliments'. *Mind*, 127(507):691–728.

_____, 2019. 'Elusive Externalism'. *Mind*, 128(510):397–427.

_____, 2020. 'The Value of Evidence'. In Maria Lasonen-Aarnio and Clayton Littlejohn, eds., *The Routledge Handbook for the Philosophy of Evidence*. Routledge.

Savage, Leonard J, 1954. *The Foundations of Statistics*. Wiley Publications in Statistics.

Schervish, Mark J, 1989. 'A general method for comparing probability assessors'. *The annals of statistics*, 17(4):1856–1879.

Schoenfield, Miriam, 2012. 'Chilling out on epistemic rationality'. *Philosophical Studies*, 158(2):197–219.

_____, 2014. 'Permission to Believe: Why Permissivism is True and What it Tells Us About Irrelevant Influences on Belief'. *Nous*, 48(2):193–218.

_____, 2016a. 'An Accuracy Based Approach to Higher Order Evidence'. *Philosophy and Phenomenological Research*, To Appear.

_____, 2016b. 'Bridging Rationality and Accuracy'. *Journal of Philosophy*, 112(12):633–657.

_____, 2017. 'Conditionalization Does Not (In General) Maximize Expected Accuracy'. *Mind*, 126(504):1155–1187.

Schultheis, Ginger, 2018. 'Living on the Edge: Against Epistemic Permissivism'. *Mind*, 127(507):863–879.

Skyrms, Brian, 1980. 'Higher Order Degrees of Belief'. In D H Mellor, ed., *Prospects for Pragmatism*, 109–137. Cambridge University Press.

_____, 1990. 'The Value of Knowledge'. *Minnesota Studies in the Philosophy of Science*, 14:245–266.

_____, 1997. 'The structure of radical probabilism'. In *Probability, dynamics and causality*, 145–157. Springer.

Stalnaker, Robert, 2019. 'Rational Reflection, and the Notorious Unmarked Clock'. In *Knowledge and Conditionals: Essays on the Structure of Inquiry*, 99–112. Oxford University Press.

Titelbaum, Michael, 2015. 'Rationality's Fixed Point (or: In Defense of Right Reason)'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 5, 253–292. Oxford University Press.

van Fraassen, Bas, 1984. 'Belief and the Will'. *The Journal of Philosophy*, 81(5):235–256.

White, Roger, 2005. 'Epistemic Permissiveness'. *Philosophical Perspectives*, 445–459.

Williamson, Timothy, 2000. *Knowledge and its Limits*. Oxford University Press.

———, 2008. 'Why Epistemology Cannot be Operationalized'. In Quentin Smith, ed., *Epistemology: New Essays*, 277–300. Oxford University Press.
———, 2014. 'Very Improbable Knowing'. *Erkenntnis*, 79(5):971–999.
———, 2018. 'Evidence of evidence in epistemic logic'. *Higher-Order Evidence: New Essays, volume To appear*.
———, 2019. 'Evidence of Evidence in Epistemic Logic'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 265–297. Oxford University Press.

How to cite this article: Dorst K, Levinstein BA, Salow B, et al. Deference Done Better. *Philosophical Perspectives*. 2021;35:99–150. https://doi.org/10.1111/phpe.12156

APPENDIX A: GLOSSARY

Definitions of technical terms and symbols are repeated here, in alphabetical order:

- **Biconvex set**: A set of points in $B \subseteq \mathbb{R}^n$ is *biconvex* iff both it and its complement $\mathbb{R}^n \setminus B$ are convex.
- $\boldsymbol{C_\pi, C_i, C_\pi^-}$: C_π is the set of candidates that π leaves open might be the expert: $C_\pi := \{\rho : \pi(P = \rho) > 0\}$. For $\mathcal{P}_i$ that occur in a frame, C_i abbreviates $C_{\mathcal{P}_i}$. $C_\pi^- := C_\pi \setminus \{\pi\}$ are the candidates it leaves open other than itself.
- **Convex**: A set of points $C \subseteq \mathbb{R}^n$ is *convex* iff it contains any average of the points it contains: if $c_1, \dots, c_n \in C$, then for any $\lambda_i \geq 0$ that sum to 1, $\sum_i \lambda_i c_i \in C$.
- **Convex hull**, ***CH***: The *convex hull* of a set of points $\rho_1, \dots, \rho_n$ is the set of points that can be obtained by taking averages of them: $CH(\{\rho_1, \dots, \rho_n\}) = \{\rho : \exists \lambda_i \geq 0 \text{ and } \sum_i \lambda_i = 1 \text{ such that } \rho = \sum_i \lambda_i \rho_i\}$.
- **Cut**: A *cut* through probability space $\mathbb{R}^n$ is a hyperplane, i.e. a set of the form $\{\pi : \mathbb{E}_\pi(X) = t\}$ for some random variable X and threshold t.
- **Decision problem**, $\mathcal{O}$: A finite set of options $\mathcal{O}$ which are functions from worlds w to numbers (utilities) $O(w)$.
- **Estimate-inaccuracy measure**, $\boldsymbol{I_X(e), I_X(\pi), I_X(P)}$: Given a random variable X, an estimate-inaccuracy measure I_X takes an estimate $e \in \mathbb{R}$ and a world w and outputs the inaccuracy of e at w, $I_X(e, w) \in \mathbb{R}$. '$I_X(\pi)$' abbreviates the inaccuracy of π's estimate: $I_X(\pi) = I_X(\mathbb{E}_\pi(X))$. '$I_X(P)$' is a definite description for the inaccuracy of the expert's estimate, whatever it is: $I_X(P)(w) := I_X(\mathbb{E}_w(X), w)$.
- **Epistemic Value**: π *epistemically values* a frame iff for any random variable X and generally strictly proper estimate-inaccuracy measure I_X, $\mathbb{E}_\pi(I_X(P)) \leq \mathbb{E}_\pi(I_X(\pi))$, with equality if and only if $\pi(\mathbb{E}(X) = \mathbb{E}_\pi(X)) = 1$.
- **Expected value**, $\boldsymbol{\mathbb{E}_\pi(X), \mathbb{E}_w, \mathbb{E}(X), \mathbb{E}_\pi(S)}$: the expected value (estimate) of a random variable X relative to π is $\mathbb{E}_\pi(X) = \sum_w \pi(w)X(w)$. For probability functions $\mathcal{P}_w$ in the frame we abbreviate $\mathbb{E}_{\mathcal{P}_w}$ to $\mathbb{E}_w$. '$\mathbb{E}(X)$' is a definite description for the expert's estimate of X, whatever it is. $\mathbb{E}_\pi(S) := \sum_w \pi(w)S_w(w)$ is the expected utility of following strategy S.
- **Fixed-option Dutch book**: Given a π and frame $\langle W, \mathcal{P}\rangle$, a fixed-option Dutch book is a pair of decision problems $\mathcal{O}_1$ and $\mathcal{O}_2$, each of which contain a "no bet" 0-option, such that O maximizes expected value amongst $\mathcal{O}_1$ relative to π, S is recommended for $\mathcal{O}_2$ by the frame, and for all $w \in W$, $O(w) + S_w(w) < 0$.

- **Generally strictly proper (gsp)**: An estimate-inaccuracy measure I_X is gsp iff for any probabilistic π, $\mathbb{E}_\pi(I_X(\pi)) \leq \mathbb{E}_\pi(I_X(\rho))$ with equality only if $\mathbb{E}_\pi(X) = \mathbb{E}_\rho(X)$.
- **Indicator variable**, $\mathbb{1}_q$: A random variable such that $\mathbb{1}_q(w) = 1$ if $w \in q$ and $\mathbb{1}_q(w) = 0$ if $w \notin q$.
- **Informed expert**, $\widehat{\boldsymbol{P}}, \widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{w}}$: The opinions the expert would have were they informed that they were the expert: $\widehat{\mathcal{P}}_w := \mathcal{P}_w(\cdot|P = \mathcal{P}_w)$. '$\widehat{P}$' is a definite description for the informed expert opinions, whatever they are.
- **Modestly informed**: A candidate $\mathcal{P}_i$ is *modestly informed* iff it's in the convex hull of $\{\widehat{\mathcal{P}}_i\} \cup C_i^-$, iff $\mathcal{P}_i = \lambda_{ii}\widehat{\mathcal{P}}_i + \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathcal{P}_j$ for $\lambda_{ij} \geq 0$ such that $\sum_j \lambda_{ij} = 1$.
- **New Reflection**: π *new-reflects* a frame iff for every function ρ, $\pi(\cdot|P = \rho) = \rho(\cdot|P = \rho)$. Equivalently (informed version): $\pi(\cdot|\widehat{P} = \rho) = \rho$.
- **Option**, $\boldsymbol{O}$: An option O is a function from worlds w to real numbers $O(w)$ represented the utility that would be achieved by taking option O at w.
- '$\boldsymbol{P}$', '$\boldsymbol{\mathcal{P}_w}$', '$\boldsymbol{\pi}$', and '$\boldsymbol{\rho}$': '$P$' is a definite description for the expert credence function, whatever it is. '$\mathcal{P}_w$' is a rigid designator for the credence function the expert has at world w. 'π' (along with other lower-case Greek letters, like 'ρ') is a rigid designator for an arbitrary probability function.
- $[\boldsymbol{P} \in \boldsymbol{C}] := \{w \in W : \mathcal{P}_w \in C\}$ is the proposition (set of worlds in a probability frame $\langle W, \mathcal{P}\rangle$) that the expert's credence function is in the set C.
- **Probability frame** $\langle \boldsymbol{W}, \boldsymbol{\mathcal{P}}\rangle$: A finite set of worlds W and a function $\mathcal{P}$ from worlds $w \in W$ to probability distributions $\mathcal{P}_w$ defined over the subsets of W, thought of as the expert's credences at w.
- **Random variable**, $\boldsymbol{X}$: A random variable X is any function from worlds w to real numbers $X(w)$.
- **Reflection**: π *reflects* a frame iff for every function ρ, $\pi(\cdot|P = \rho) = \rho$. Equivalently, iff for every convex set C: $\pi(\cdot|P \in C) \in C$.
- **Simple Trust**: π *simply trusts* a frame iff, for all q, t: $\pi(q|P(q) \geq t) \geq t$.
- **Strategy**, $\boldsymbol{S}$: given a decision problem $\mathcal{O}$, a *strategy* is a function from worlds w to options $S_w \in \mathcal{O}$ such that $S_w = S_v$ whenever $\mathcal{P}_w = \mathcal{P}_v$. S is *recommended for* $\mathcal{O}$ by a frame iff, for each $w \in W$, S_w maximizes expected utility relative to $\mathcal{P}_w$ amongst the options: $\mathbb{E}_w(S_w) \geq \mathbb{E}_w(O)$ for any $O \in \mathcal{O}$.
- **Total Trust**: π *totally trusts* a frame iff for any random variable X and threshold $t \in \mathbb{R}$, $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) \geq t$. Equivalently, for every biconvex set B, $\pi(\cdot|P \in B) \in B$.
- **Trust**: π *trusts* a frame iff for any q, p, t: $\pi(q|p \wedge [P(q|p) \geq t]) \geq t$.
- **Validates**: A frame $\langle W, \mathcal{P}\rangle$ validates a deference principle Φ iff for every $i \in W$, $\mathcal{P}_i$ defers Φ-wise to the frame.
- **Value**: π *values* a frame $\langle W, \mathcal{P}\rangle$ iff for any decision problem $\mathcal{O}$, any recommended strategy has higher expected utility than any option: if S is recommended for $\mathcal{O}$ by the frame, then $\mathbb{E}_\pi(S) = \sum_v \pi(v)S_v(v) \geq \mathbb{E}_\pi(O)$ for any $O \in \mathcal{O}$.
- $\boldsymbol{W_\pi} : \{w \in W : \pi(w) > 0\}$.

APPENDIX B: PROOFS

For both efficiency and technical reasons we will prove the main results using a slightly different structure than the theorems stated in the text. Rather than proving a series of biconditionals, will will first prove a cycle, and then use all of these results together for the final link. In particular, for technical reasons it is much harder to prove the full version of Value directly; instead, we first work with a potentially weaker version (which we'll later show to in fact be equivalent):

Weak Value: Given any $\mathcal{O}$, there *exists* some recommended strategy S such that for all $O \in \mathcal{O}$: $\mathbb{E}_\pi(S) \geq \mathbb{E}_\pi(O)$.

The difference is that Value universally quantifies over recommended strategies, whereas Weak Value existentially quantifies over them. Note that for frames-plus-decision-problems for which the recommended strategy is unique, Weak Value holds iff Value does.

We'll proceed as follows. We'll first (§B.1) prove that π weakly values a frame only if π totally trusts it. We'll then prove that π totally trusts a frame only if π is in the convex hull of C_π and each $\mathcal{P}_i \in C_\pi$ is modestly informed. We'll then show that if this condition holds, then π weakly values the frame. This shows that these three conditions are equivalent. Next (§B.2), we'll show that, together, they entail that π values the frame—thus establishing Theorems 2.2 and 4.1. Finally, in §B.3 we'll prove Theorem 3.2—that Total Trust is equivalent to epistemic value—directly. Together, these results establish Theorem 5.1.

B.1 | Weak Value ⇔ Total Trust ⇔ Modestly Informed

Lemma B.1. *If π weakly values $\langle W, \mathcal{P}\rangle$, π totally trusts it.*

Proof. Supposing π doesn't totally trust $\langle W, \mathcal{P}\rangle$, we find a decision problem using the same random variable on which Value fails and in which there is a unique recommended strategy—therefore, Weak Value fails too.

If π doesn't totally trust the frame, there is an X, t such that $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) = a < t$. (Note that this implies $\pi(\mathbb{E}(X) \geq t) > 0$.) Find the maximum $b < t$ such that $\exists w \in W$: $\mathbb{E}_w(X) = b$, and let s be any number strictly between $\max(a, b)$ and t. Let Y be a random variable that takes values s at all worlds, and let $\mathcal{O} = \{X, Y\}$.

By construction, for all $x \in [\mathbb{E}(X) \geq t]$, $\mathbb{E}_x(X) > \mathbb{E}_x(Y)$; and for all $y \in [\mathbb{E}(X) < t]$, $\mathbb{E}_y(X) < \mathbb{E}_y(Y)$. Thus there is a uniquely recommended strategy S—namely, $S_w = X$ iff $w \in [\mathbb{E}(X) \geq t]$ and $S_w = Y$ iff $w \in [\mathbb{E}(X) < t]$. Noting that $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) < s$ and that $\mathbb{E}_\pi(Y|\mathbb{E}(X) < t) = s$ if defined, we then have:

$$\begin{aligned}
\mathbb{E}_\pi(S) &= \pi(\mathbb{E}(X) \geq t) \cdot \mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) \quad + \pi(\mathbb{E}(X) < t) \cdot \mathbb{E}_\pi(Y|\mathbb{E}(X) < t) \\
&< \pi(\mathbb{E}(X) \geq t \cdot s \quad + \pi(\mathbb{E}(X) < t) \cdot s \\
&= s = \mathbb{E}_\pi(Y).
\end{aligned}$$

We thus have $\mathbb{E}_\pi(Y) > \mathbb{E}_\pi(S)$; Weak Value fails. □

The next step is to prove:

Lemma B.2. *If π totally trusts $\langle W, \mathcal{P}\rangle$, then π is in the convex hull of C_π and each $\mathcal{P}_i \in C_\pi$ is modestly informed.*

To do so, we first prove some lemmas about the properties of C_π if all the candidates are modestly informed; in particular, we want to show that that assuming π is in the convex hull of C_π, all the candidates are modestly informed iff a more general condition (class-convexity, defined below) holds.

Remark B.3. When $\pi = \sum_i \lambda_i \mathcal{P}_i$, or $\mathcal{P}_j = \lambda_{jj}\widehat{\mathcal{P}}_j + \sum_{\mathcal{P}_i \in C_i^-} \lambda_{ji}\mathcal{P}_i$, there may well be multiple worlds $w \neq w'$ such that $\mathcal{P}_w = \mathcal{P}_i = \mathcal{P}_{w'}$. Nevertheless, we can always choose a single representative world i in each such equivalence class, and in what follows we write the terms λ_{ji} (etc.) assuming we have done so.

Lemma B.4. *Take any nonempty set of points $A = \{\mathcal{P}_1, \dots, \mathcal{P}_n\}$ that are each modestly informed, so $\mathcal{P}_i = \lambda_{ii}\widehat{\mathcal{P}}_i + \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathcal{P}_j$, and which assign weight only to each other, so $\lambda_{ij} > 0$ only if $\mathcal{P}_j \in A$. Then $\lambda_{ii} > 0$ for some $\mathcal{P}_i$ in A.*

Proof. Suppose not: each $\mathcal{P}_i = \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathcal{P}_j$ where each $C_i^- \subset \{\mathcal{P}_1, \dots, \mathcal{P}_n\}$. Then each $\mathcal{P}_i$ is in the convex hull of the other $\mathcal{P}_j$, meaning that $CH(A)$ has at most one extreme point and therefore is a singleton. But since $\mathcal{P}_1 \notin C_1^-$, $\mathcal{P}_1 = \sum_{\mathcal{P}_j \in C_1^-} \lambda_{1j}\mathcal{P}_j = 0$, contradicting the fact that $\mathcal{P}_1$ is a probability function. □

Definition B.5. Let $\boldsymbol{W_\pi}$ be the set of worlds seen by π: $\{w \in W : \pi(w) > 0\}$.

Lemma B.6 (Transitivity). *If each $\mathcal{P}_i$ in C_π is modestly informed and π is in their convex hull, then they all are such that $\mathcal{P}_i(W_\pi) = 1$*

Proof. Let $W'_\pi := \{w \in W_\pi : \mathcal{P}_w(W_\pi) = 1\}$. W'_π is nonempty, otherwise π is not in C_π's convex hull. We first note that every $\mathcal{P}_i$ for $i \in W'_\pi$, is a mixture of $\widehat{\mathcal{P}}_i$ and $\{\mathcal{P}_j : j \in W'_\pi\}$. For $\mathcal{P}_i = \lambda_{ii}\widehat{\mathcal{P}}_i + \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathcal{P}_j$, and since $\mathcal{P}_i(W_\pi) = 1$, $C_i \subseteq \{\mathcal{P}_j : j \in W_\pi\}$; yet for any $\mathcal{P}_j$ with $j \in W_\pi$ but not in W'_π, $\mathcal{P}_j(W_\pi) < 1$, so $\mathcal{P}_j(x) > 0$ for $x \notin W_\pi$; hence if $\lambda_{ij} > 0$, then $\mathcal{P}_i(x) \geq \lambda_{ij}\mathcal{P}_j(x) > 0$, contradicting the assumption that $\mathcal{P}_i(W_\pi) = 1$.

Next we show that for $i \in W'_\pi$, $\mathcal{P}_i(W'_\pi) = 1$. Suppose not, so $\mathcal{P}_i(x) > 0$ for $x \in W_\pi$ but not in W'_π. Let $t := \max_{w \in W'_\pi}(\mathcal{P}_w(x))$, and let $M := \{w \in W'_\pi : \mathcal{P}_w(x) = t\}$ and $m \in M$. Note that (by the above) $\lambda_{mj} > 0$ only if $j \in W'_\pi$; hence $\lambda_{mj} > 0$ only if $\mathcal{P}_j(x) \leq t$. Thus if $\lambda_{mk} > 0$ for $k \notin M$, $\mathcal{P}_j(x)$ would average to less than t; so $\lambda_{mj} > 0$ only if $j \in M$. Notice also that since $\widehat{\mathcal{P}}_m(x) = 0$ (since $\mathcal{P}_x(W_\pi) < 1$ but $\mathcal{P}_m(W_\pi) = 1$), we must similarly have that $\lambda_{mm} = 0$, for $\mathcal{P}_m(x) \leq \lambda_{mm} \cdot 0 + (1 - \lambda_{mm})t$. Hence we have that for all $m \in M$: $\lambda_{mm} = 0$ but $\lambda_{mj} > 0$ only if $j \in M$; i.e. each such m is modestly informed, assign weight only to each other, and assign no weight to themselves. By Lemma B.4, this is a contradiction; so we have that $\mathcal{P}_i(W'_\pi) = 1$.

We can now show that $W_\pi \setminus W'_\pi = \emptyset$. For we know that $\pi = \sum_i \lambda_i \mathcal{P}_i$ for $\mathcal{P}_i \in C_\pi$. Since $\pi(W_\pi) = 1$ we know $\lambda_i > 0$ only if $\mathcal{P}_i(W_\pi) = 1$, hence $\lambda_i > 0$ only if $\mathcal{P}_i \in W'_\pi$. But since we now know that all such $\mathcal{P}_i$ have $\mathcal{P}_i(W'_\pi) = 1$, it follows that $\pi(W'_\pi) = 1$, i.e. $W_\pi = W'_\pi$. Hence for all $i \in W_\pi = W'_\pi$, we have $\mathcal{P}_i(W_\pi) = \mathcal{P}_i(W'_\pi) = 1$. □

Lemma B.7 (Reflexivity). *If each $\mathcal{P}_i \in C_\pi$ is modestly informed and π is in their convex hull, then for all $i \in W_\pi$, $\mathcal{P}_i(i) > 0$.*

Proof. Suppose $\pi(i) > 0$ but $\mathcal{P}_i(i) = 0$. Since $\pi = \sum_{\mathcal{P}_j \in C_\pi} \lambda_j \mathcal{P}_j$, there must be some $j \in W_\pi$ such that $\mathcal{P}_j(i) > 0$. Let $t = \max_{j \in W_\pi}(\mathcal{P}_j(i))$ and $M := \{j \in W_\pi : \mathcal{P}_j(i) = t\}$, and $m \in M$. By Lemma B.6, $\mathcal{P}_m(W_\pi) = 1$, so $\lambda_{mk} > 0$ only if $k \in W_\pi$, so only if $\mathcal{P}_k(i) \leq t$. Thus if $\lambda_{mk} > 0$ for $k \notin M$, $\mathcal{P}_m(i)$ must average to less than t; but it doesn't. Similarly, note that since $\widehat{\mathcal{P}}_m(i) = 0$ (since $\mathcal{P}_i \neq \mathcal{P}_m$), we must likewise have that $\lambda_{mm} = 0$. So M is a nonempty set of worlds which assign

weight $\lambda_{mk} > 0$ only to each other and not to themselves—contradicting Lemma B.4. Hence if $i \in W_\pi$, then $\mathcal{P}_i(i) > 0$. □

Definition B.8. W_π is **class-convex** iff each candidate $\mathcal{P}_i$ is in the convex hull of its informed self and the other candidates in W_π (as opposed to the other candidates *it* leaves open, for modest-informedness): for all $\mathcal{P}_i \in C_\pi$, $\mathcal{P}_i = \lambda_{ii}\widehat{\mathcal{P}}_i + \sum_{\mathcal{P}_j \neq \mathcal{P}_i : \mathcal{P}_j \in C_\pi} \lambda_{ij}\mathcal{P}_j$.

Lemma B.9 (Class-Convexity). *If π is in the convex hull of C_π, then W_π is class-convex iff each $\mathcal{P}_i \in C_\pi$ is modestly informed.*

Proof. ($\Rightarrow$) : Suppose W_π is class-convex but there is a $\mathcal{P}_i \in C_\pi$ that's not modestly informed. Since W_π is class-convex, $\mathcal{P}_i$ *is* in the convex hull of its informed self and the other candidates in W_π: $\mathcal{P}_i = \lambda_{ii}\widehat{\mathcal{P}}_i + \sum_{\mathcal{P}_j \neq \mathcal{P}_i : \mathcal{P}_j \in C_\pi} \lambda_{ij}\mathcal{P}_j$. By Lemma B.6, $\mathcal{P}_i(W_\pi) = 1$, so $C_i \subseteq C_\pi$. Since $\mathcal{P}_i$ is *not* in the convex hull of $\widehat{\mathcal{P}}_i$ and C_i^-, this means there is some $\mathcal{P}_j \in C_\pi$ but $\mathcal{P}_j \notin C_i$ such that $\lambda_{ij} > 0$. By Lemma B.7, $\mathcal{P}_j(j) > 0$, hence $\mathcal{P}_i(j) \geq \lambda_{ij}\mathcal{P}_j > 0$. But since $\mathcal{P}_j \notin C_i$, $\mathcal{P}_i(j) = 0$—contradiction.

($\Leftarrow$) : If W_π is not class-convex, then there is a $\mathcal{P}_i$ that's not in the convex hull of $\widehat{\mathcal{P}}_i$ and $\{\mathcal{P}_j \neq \mathcal{P}_i : \mathcal{P}_j \in C_\pi\}$. Since by Lemma B.6 $C_i \subseteq C_\pi$, it follows that $\mathcal{P}_i$ is not in the convex hull of $\widehat{\mathcal{P}}_i$ and C_i^-, so it not modestly informed. □

We're now in a position to prove Lemma B.2, repeated here:

Lemma B.10. *If π totally trusts $\langle W, \mathcal{P}\rangle$, then π is in the convex hull of C_π and each $\mathcal{P}_i \in C_\pi$ is modestly informed.*

Proof. Suppose π is not in the convex hull of C_π. Then by the hyperplane separation theorem, there is an X, t strongly separating them: i.e., $\mathbb{E}_\pi(X) < t$ but for all $\mathcal{P}_i \in C_\pi$, $\mathbb{E}_i(X) \geq t$. Then $\pi(\mathbb{E}(X) \geq t) = 1$, so $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) = \mathbb{E}_\pi(X) < t$; Total Trust fails.

Now suppose that π is in the convex hull of C_π yet some $\mathcal{P}_i \in C_\pi$ is not modestly informed. By Lemma B.9, this means W_π is not class-convex, so some $\mathcal{P}_j$ is not in the convex hull of $\widehat{\mathcal{P}}_j$ and $\{\mathcal{P}_k \neq \mathcal{P}_j : \mathcal{P}_k \in C_\pi\}$. By the hyperplane separation theorem, there must be an X, t that strongly separates them: $\mathbb{E}_j(X) \geq t$, but $\widehat{\mathbb{E}}_j(X) < t$ and likewise for all $\mathcal{P}_k \neq \mathcal{P}_i$ in C_π, $\mathbb{E}_k(X) < t$. By the latter fact, $[\mathbb{E}(X) \geq t] = [P = \mathcal{P}_j]$. If Total Trust held, we'd have $\mathbb{E}_\pi(X|\mathbb{E}(X) \geq t) = \mathbb{E}_\pi(X|P = \mathcal{P}_j) = \widehat{\mathbb{E}}_j(X) \geq t$ (since, by endnotes 36 and 37, Total Trust entails New Reflection); but by the above, $\widehat{\mathbb{E}}_j(X) < t$; so Total Trust fails. □

Next, we prove the last link in this cycle:

Lemma B.11. *If π is in the convex hull of C_π and each $\mathcal{P}_i \in C_\pi$ is modestly informed, then π weakly values $\langle W, \mathcal{P}\rangle$.*

Proof. By Lemma B.9, we know that W_π is class convex, and by Lemma B.6, each $\mathcal{P}_i \in C_\pi$ has $\mathcal{P}_i(W_\pi) = 1$. Thus throughout we restrict quantification over worlds to those in W_π.

We first prove that each $\mathcal{P}_i$ in C_π weakly values the frame, and moreover that there is always a single strategy that they all value. Suppose not, so there is a decision problem $\mathcal{O}$ such that for any recommended strategy S, there is a world i and an option $O \in \mathcal{O}$ such that $\mathbb{E}_i(O) > \mathbb{E}_i(S)$, hence $\mathbb{E}_i(O - S) > 0$. For any j, let $\mathcal{M}_j := \{O : (\forall O')\mathbb{E}_j(O) \geq \mathbb{E}_j(O')\}$ be the set of options with maximal

(uninformed) expected utility at world j; so any S such that for all $j \in W, S_j \in \mathcal{M}_j$ will be a recommended strategy. Choose a strategy S such that for all $j \in W$: S_j maximizes *informed* expected utility amongst the options with maximal (uninformed) expected utility: $\forall O' \in \mathcal{M}_j, \widehat{\mathbb{E}}_j(S_j) \geq \widehat{\mathbb{E}}_j(O')$. This is a recommended strategy, and so still we have some i, O such that $\mathbb{E}_i(O - S) > 0$. We use this fact to show that class-convexity must fail, contradicting our hypothesis.

Find a pair $\langle O', m\rangle$ that maximizes this divergence in the frame: $\mathbb{E}_m(O' - S) = \alpha = \max_{\langle O'', j\rangle}(\mathbb{E}_j(O'' - S))$. (It exists, since the number of options and worlds are both finite.) Let $M := \{j \in W : \mathbb{E}_j(O' - S) = \alpha\}$ be the set of worlds with this maximal divergence. Note three facts about M:

F1: If $k \notin M$, then $\mathbb{E}_k(O' - S) < \alpha$, by construction.
F2: If $j \in M$, then O' has maximal expected value amongst $\mathcal{O}$. (If not, then α is not the maximal divergence after all, for there is an O'' such that $\mathbb{E}_j(O'') > \mathbb{E}_j(O')$, and therefore $\mathbb{E}_j(O'' - S) > \mathbb{E}_j(O' - S) = \alpha$.)
F3: If $j \in M$, then $\widehat{\mathbb{E}}_j(O' - S) \leq 0 < \alpha$, since O' maximizes $\mathbb{E}_j$, and S_j has maximal informed expected value amongst such options, so $\widehat{\mathbb{E}}_j(S_j) \geq \widehat{\mathbb{E}}_j(O')$. Since $\widehat{\mathcal{P}}_j(S = S_j) = 1$ since $\widehat{\mathcal{P}}_j$ knows what P is, it follows that $\widehat{\mathbb{E}}_j(S) = \widehat{\mathbb{E}}_j(S_j) \geq \widehat{\mathbb{E}}_j(O')$, so $\widehat{\mathbb{E}}_j(O' - S) \leq 0$.

Now let $A := CH(\{\mathcal{P}_k : k \notin M\} \cup \{\widehat{\mathcal{P}}_j : j \in M\})$ be the convex hull of the uninformed opinions outside M and the informed opinions inside it. By F1 and F3, $\forall \rho \in A$: we have that $\mathbb{E}_\rho(O' - S) < \alpha$, while for all $j \in M$: $\mathbb{E}_j(O' - S) \geq \alpha$. Thus $\langle O' - S, \alpha\rangle$ determines a hyperplane that separates all the $\mathcal{P}_j$ (for $j \in M$) from A, meaning $\mathcal{P}_j$ is not in A.

We now strengthen this conclusion to show that there must be a $\mathcal{P}_i$ for $i \in M$ on which class-convexity fails, i.e. $\mathcal{P}_i \notin A_i := CH(\{\widehat{\mathcal{P}}_i\} \cup \{\mathcal{P}_j \neq \mathcal{P}_i : \mathcal{P}_j \in C_\pi\})$. Note that

$$\begin{aligned} A_i \subseteq A_i^* &:= CH(\{\widehat{\mathcal{P}}_j : j \in M\} \cup \{\mathcal{P}_k : k \notin M\} \cup \{\mathcal{P}_j \neq \mathcal{P}_i : j \in M\}) \\ &= CH(A \cup \{\mathcal{P}_j \neq \mathcal{P}_i : j \in M\}) \end{aligned}$$

so it'll suffice to show that $\mathcal{P}_i$ is separable from A_i^*.

Take a $\mathcal{P}_i$ that is extreme within the convex hull of $\{\mathcal{P}_j : j \in M\}$, so $\mathcal{P}_i$ is not in the convex hull of $\{\mathcal{P}_j \neq \mathcal{P}_i : j \in M\}$. Suppose, for reductio, that $\mathcal{P}_i$ *is* in the convex hull of A_i^*, so there are $\rho_k \in A$ such that $\mathcal{P}_i = \sum_k \lambda_k \rho_k + \sum_{\mathcal{P}_j \neq \mathcal{P}_i : j \in M} \lambda_j \mathcal{P}_j$. Now, if $\lambda_k = 0$ for all ρ_k, then $\mathcal{P}_i$ would be in the convex hull of $\{\mathcal{P}_j \neq \mathcal{P}_i : j \in M\}$, contradicting the assumption that it's extreme within M; so $\lambda_k > 0$ for some ρ_k. But we know that $\mathbb{E}_\rho(O' - S) \leq \alpha - \varepsilon$ for $\varepsilon > 0$, while $\mathbb{E}_j(O' - S) = \alpha$ for all $j \in M$; hence $\mathbb{E}_i(O' - S) \leq \lambda_k(\alpha - \varepsilon) + (1 - \lambda_k)\alpha < \alpha$, contradicting the assumption that $i \in M$. Thus $\mathcal{P}_i$ is *not* in A_i^*, and hence it is not in A_i, so W_π is not class-convex.

This establishes that for any decision problem $\mathcal{O}$, there is a strategy S such that for all $i \in W_\pi$ and $O \in \mathcal{O}$, $\mathbb{E}_i(S) \geq \mathbb{E}_i(O)$, so $\mathbb{E}_i(S - O) \geq 0$. Note that since π is in the convex hull of the $\mathcal{P}_i$ for $i \in W_\pi$, this means that for any such $\mathcal{O}$, there is an S such that for all O, $\mathbb{E}_\pi(S - O) = \sum_i \lambda_i \mathbb{E}_i(S - O) \geq \sum_i \lambda_i 0 = 0$, and hence $\mathbb{E}_\pi(S) \geq \mathbb{E}_\pi(O)$, and so π weakly values the frame. □

By Lemma B.1, π weakly values a frame only if π totally trusts it; by Lemma B.2, π totally trusts it only if π is in the convex hull of C_π and each $\mathcal{P}_i \in C_\pi$ is modestly informed; by Lemma B.11 only if π weakly values the frame. Hence we've shown that these three conditions are equivalent.

B.2 | Weak Value ⇔ Value

Having established that Weak Value is equivalent to both Total Trust and our convex-hull-plus-modestly-informed constraint, we now prove that together all these principles entail full Value. Since full Value (obviously) entails Weak Value (since there always is at least one recommended strategy), this will establish that all these conditions are equivalent.

We first show a helpful Lemma about the decomposition of modestly informed $\mathcal{P}_i$:

Lemma B.12. *If π is in the convex hull of C_π and each $\mathcal{P}_i \in C_\pi$ is modestly informed, so $\mathcal{P}_i = \lambda_{ii}\widehat{\mathcal{P}}_i + \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathcal{P}_j$, then $\lambda_{ii} > 0$.*

Proof. Note that by the results of §B.1, both π and all the $\mathcal{P}_i$ must totally trust the frame. Suppose $\lambda_{ii} = 0$, so $\mathcal{P}_i$ is in the convex hull of C_i^-. Take $\mathcal{P}_k \in \arg\max_{\mathcal{P}_j \in C_i^-}(\mathcal{P}_k(P = \mathcal{P}_i))$. Since $\mathcal{P}_i$ is in the convex hull of C_i^-, $\mathcal{P}_i(P = \mathcal{P}_i) \leq \mathcal{P}_k(P = \mathcal{P}_i)$. If there is a $w \in [P = \mathcal{P}_i]$ such that $t = \mathcal{P}_k(w) > \mathcal{P}_i(w)$, then $\pi(w|P(w) \geq t) = 0$, contradicting (Simple) Trust. Thus for all $w \in [P = \mathcal{P}_i]$, $\mathcal{P}_i(w) \geq \mathcal{P}_k(w)$, hence we have equality: $\mathcal{P}_i(P = \mathcal{P}_i) = \mathcal{P}_k(P = \mathcal{P}_i)$. Since nevertheless $\mathcal{P}_i \neq \mathcal{P}_k$, there must be an $x \notin [P = \mathcal{P}_i]$ such that $\mathcal{P}_k(x) > \mathcal{P}_i(x)$ and hence $\mathcal{P}_k(\neg\{x\}) < \mathcal{P}_i(\neg\{x\})$. Thus we have that

$$t' := \mathcal{P}_k(P = \mathcal{P}_i|\neg\{x\}) = \frac{\mathcal{P}_k(P = \mathcal{P}_i)}{\mathcal{P}_k(\neg\{x\})} > \frac{\mathcal{P}_i(P = \mathcal{P}_i)}{\mathcal{P}_i(\neg\{x\})} = \mathcal{P}_i(P = \mathcal{P}_i|\neg\{x\})$$

Thus if there's some world in W_π that's not x and where $[P(P = \mathcal{P}_i|\neg\{x\}) \geq t']$ holds, we have that $\pi(P = \mathcal{P}_i|\neg\{x\} \wedge [P(P = \mathcal{P}_i|\neg\{x\}) \geq t'])$ is well-defined and equal to $0 < t'$ (since no $w \in [P = \mathcal{P}_i]$ is such a world), contradicting the fact that π trusts the frame. Conversely, if there is no such world, that means that $k = x$ and so then $\mathcal{P}_k(P(P = \mathcal{P}_i|\neg\{x\}) < t'|\neg\{x\}) = 1$. It follows that Trust fails at $\mathcal{P}_k$, since we have that $\mathcal{P}_k(P = \mathcal{P}_i|\neg\{x\} \wedge [P(P = \mathcal{P}_i|\neg\{x\}) < t']) = \mathcal{P}_k(P = \mathcal{P}_i|\neg\{x\}) \geq t'$, contradicting the fact that $\mathcal{P}_k$ trusts the frame. □

We're now in a position to prove that Weak Value entails full Value. The strategy is to take a case where Value fails, and show that we can modify it by slightly adjusting the available options to make there be a *unique* recommended strategy, to generate a case where Weak Value fails. To do so, we will need to use our knowledge about what the frame must look like for Weak Value to hold, using the Lemmas B.1–B.12.

Lemma B.13. *π weakly values a frame iff it values it.*

Proof. The ⇐ direction is immediate, so we show the ⇒ direction. Suppose π weakly values the frame $\langle W, \mathcal{P}\rangle$. By Lemmas B.1–B.11 we know that π also totally trusts the frame and that each $\mathcal{P}_i \in C_\pi$ is modestly informed and that π is in their convex hull. Moreover, since for each $\mathcal{P}_i \in C_\pi$, by Lemma B.6 we know that $C_i \subseteq C_\pi$, we know that all $\mathcal{P}_j \in C_i$ are modestly informed; and since by Lemma $B.7$ we have that $\mathcal{P}_i(i) > 0$, $\mathcal{P}_i \in C_i$ and so $\mathcal{P}_i$ is automatically in the convex hull of C_i. Thus applying Lemmas B.1–B.11 with $\mathcal{P}_i$ substituted for π, we have that each $\mathcal{P}_i$ also weakly values and totally trusts the frame.

Suppose, for reductio, that π *doesn't* value the frame, so that there is a decision problem $\mathcal{O} = \{O_1, \ldots, O_m\}$ and a recommended strategy S such that $\mathbb{E}_\pi(O - S) > 0$ for some $O \in \mathcal{O}$. We will show that we can alter the decision problem to make one on which there is a *unique* recommended strategy, and on which π still fails to value the frame—and hence, that Weak Value fails as well

(since when there is a unique recommended strategy, Weak Value holds for that decision problem iff Value does).

Relabel things so that $\{\mathcal{P}_1, \dots, \mathcal{P}_n\} = C_\pi$, and consider the set of options they select under S: $\{S_1, \dots, S_n\} \subseteq \mathcal{O}$. We are going to remove all these options from the decision problem, and replace them with ones S_i^t that are identical to S_i except they yield an additional t value when $P = \mathcal{P}_i$. In particular, for any $t \geq 0$ and i, let

$$S_i^t(w) := \begin{cases} S_i(w) + t & \text{if } \mathcal{P}_w = \mathcal{P}_i \\ S_i(w) & \text{otherwise} \end{cases}$$

Since C_π is all modestly informed, for all $\mathcal{P}_i \in C_\pi$ we have that

$$\mathbb{E}_i\left(S_i^t - S_i\right) = \lambda_{ii}\widehat{\mathbb{E}}_i\left(S_i^t - S_i\right) + \sum_{\mathcal{P}_j \in C_i^-} \lambda_{ij}\mathbb{E}_j\left(S_i^t - S_i\right)$$

where $\widehat{\mathbb{E}}_i(S_i^t - S_i) = t$ and $\mathbb{E}_j(S_i^t - S_i) \geq 0$ by definition of S_i^t, and $\lambda_{ii} > 0$ by Lemma B.12. Viewing this equation as a function of t, notice that it is continuous and monotonically increasing in t; moreover, when $t = 0$, $\mathbb{E}_i(S_i^t - S_i) = 0$, and as $t \to \infty$, $\lambda_{ii}t \to \infty$ and thus $\mathbb{E}_i(S_i^t - S_i) \to \infty$. As a result, for any $\alpha \geq 0$, by the intermediate value theorem there is a unique t_i such that $\mathbb{E}_i(S_i^{t_i} - S_i) = \alpha$. (Note that as $\alpha \to 0$, likewise $t_i \to 0$.)

Consider an arbitrary $\alpha > 0$ and for each $\mathcal{P}_i$ choose $t_i > 0$ such that $\mathbb{E}_i(S_i^{t_i} - S_i) = \alpha$. Now consider a new decision-problem $\mathcal{O}^* := (\mathcal{O} - \{S_1, \dots, S_n\}) \cup \{S_1^{t_1}, \dots, S_n^{t_n}\}$. We first show that for any $\alpha > 0$, there is a uniquely recommended strategy for $\mathcal{O}^*$; we then show that if α is small enough, this will be a decision-problem on which Value (and hence Weak Value) fails.

We know that for each $\mathcal{P}_i$, $\mathbb{E}_i(S_i)$ was maximal amongst the options in $\mathcal{O}$. Letting $\mathcal{X} = \mathcal{O} - \{S_1, \dots, S_n\}$, note that $\mathcal{O}^* = \mathcal{X} \cup \{S_1^{t_1}, \dots, S_n^{t_n}\}$. For any $O \in \mathcal{X}$, we know that $\mathbb{E}_i(S_i^{t_i}) > \mathbb{E}_i(S_i) \geq \mathbb{E}_i(O)$. So to show that each $S_i^{t_i}$ is the unique option that maximizes expected utility according to $\mathcal{P}_i$, it suffices to show that if $\mathcal{P}_i \neq \mathcal{P}_j$, then $\mathbb{E}_i(S_i^{t_i}) > \mathbb{E}_i(S_j^{t_j})$. Notice that this holds iff $\mathbb{E}_i(S_i^{t_i} - S_i) > \mathbb{E}_i(S_j^{t_j} - S_i)$, and since $\mathbb{E}_i(S_i) \geq \mathbb{E}_i(S_j)$, it suffices to show that

$$\mathbb{E}_i\left(S_i^{t_i} - S_i\right) > \mathbb{E}_i\left(S_j^{t_j} - S_j\right) \tag{$*$}$$

Note that since $\mathbb{E}_i(S_i^{t_i} - S_i) = \alpha = \mathbb{E}_j(S_j^{t_j} - S_j)$, it suffices to show that

$$\mathbb{E}_j\left(S_j^{t_j} - S_j\right) > \mathbb{E}_i\left(S_j^{t_j} - S_j\right) \tag{$**$}$$

First note that for all $\mathcal{P}_k \neq \mathcal{P}_j$, we have that $\mathbb{E}_k(S_j^{t_j} - S_j) \leq \mathbb{E}_j(S_j^{t_j} - S_j)$. For suppose not, and instead $\mathbb{E}_k(S_j^{t_j} - S_j) \geq s > \mathbb{E}_j(S_j^{t_j} - S_j) > 0$. Then $[\mathbb{E}(S_j^{t_j} - S_j) \geq s]$ is nonempty and entails $P \neq \mathcal{P}_j$, therefore by definition of $S_j^{t_j}$, $[\mathbb{E}(S_j^{t_j} - S_j) \geq s] \subseteq [S_j^{t_j} - S_j = 0]$, hence $\mathbb{E}_\pi(S_j^{t_j} - S_j | \mathbb{E}(S_j^{t_j} - S_j) \geq s) = 0 < s$, violating Total Trust. Thus $\mathbb{E}_k(S_j^{t_j} - S_j) \leq \mathbb{E}_j(S_j^{t_j} - S_j)$ for

all $\mathcal{P}_k$. Given this, for $\mathcal{P}_j \neq \mathcal{P}_i$ we have that

$$\begin{aligned}
\mathbb{E}_i\left(S_j^{t_j} - S_j\right) &= \lambda_{ii}\widehat{\mathbb{E}}_i\left(S_j^{t_j} - S_j\right) + \sum_{\mathcal{P}_k \in C_i^-} \lambda_{ik}\mathbb{E}_k\left(S_j^{t_j} - S_j\right) \\
&\leq \lambda_{ii}0 + \sum_{\mathcal{P}_k \in C_i^-} \lambda_{ik}\mathbb{E}_j\left(S_j^{t_j} - S_j\right) \\
&= (1 - \lambda_{ii})\mathbb{E}_j\left(S_j^{t_j} - S_j\right) \\
&< \mathbb{E}_j\left(S_j^{t_j} - S_j\right)
\end{aligned}$$

The last line follows from the fact that, by Lemma B.12, $\lambda_{ii} > 0$. This establishes (**), and therefore (*), and therefore that for all $\mathcal{P}_i$, $S_i^{t_i}$ is an option that uniquely maximizes expected value, i.e. that the strategy S^* such that $S_i^* = S_i^{t_i}$, for each $\mathcal{P}_i$, is the uniquely recommended strategy for $\mathcal{O}^*$, for an arbitrary $\alpha > 0$.

From here we show that Weak Value fails. We know that $\mathbb{E}_\pi(O - S) > 0$; say it equals $\beta > 0$. We know moreover that no matter which $\alpha > 0$ we choose in modifying $\mathcal{O}$ to $\mathcal{O}^*$, we'll have an option $O^* \in \mathcal{O}^*$ such that $\mathbb{E}_\pi(O^*) \geq \mathbb{E}_\pi(O)$ (since our modifications only replace options with more valuable ones). Finally, note that for any $w \in W_\pi$, there's a $\mathcal{P}_i$ such that $\mathcal{P}_w = \mathcal{P}_i$, so $(S^* - S)(w) = (S_i^{t_i} - S_i)(w) = t_i$. Hence this divergence $S^* - S$ is upper-bounded across all worlds by the maximal t_i used to modify the options S_i to $S_i^{t_i}$. Recalling that as $\alpha \to 0$, all such $t_i \to 0$, we can choose an $\alpha > 0$ small enough so that $t_1, \ldots, t_n < \beta$, in which case we have the $\mathbb{E}_\pi(S^* - S) < \beta$. It follows that

$$\begin{aligned}
\mathbb{E}_\pi(O - S) &= \beta > 0 \\
\Rightarrow \mathbb{E}_\pi(O - S) - \mathbb{E}_\pi(S^* - S) &> 0 \\
\Rightarrow \mathbb{E}_\pi(O - S - S^* + S) &> 0 \\
\Rightarrow \mathbb{E}_\pi(O - S^*) &> 0
\end{aligned}$$

And, since $\mathbb{E}_\pi(O^*) \geq \mathbb{E}_\pi(O)$, we have that $\mathbb{E}_\pi(O^* - S^*) > 0$, i.e. $\mathbb{E}_\pi(O^*) > \mathbb{E}_\pi(S^*)$, which is just to say that π does not value the frame on this decision-problem $\mathcal{O}^*$. Since S^* is the uniquely recommended strategy on this decision problem, it follows that π doesn't weakly value the frame, completing the proof. □

Combining Lemmas B.1, B.2, B.11, and B.13, we've now established:

Theorem B.14. *The following are equivalent:*

- *π values $\langle W, \mathcal{P}\rangle$.*
- *π weakly values $\langle W, \mathcal{P}\rangle$.*
- *π totally trusts $\langle W, \mathcal{P}\rangle$.*
- *π is in the convex hull of C_π an each $\mathcal{P}_i \in C_\pi$ is modestly informed.*

What remains to be done in order to establish our full characterization result, Theorem 5.1, is to prove Theorem 3.2, which we do in the next subsection.

B.3 | Accuracy theorem

We give two different proofs of Theorem 3.2 to establish the connection between Total Trust and Epistemic Value. The second (§B.3.2) is the one we started with, using Campbell-Moore's (2020) characterization of gsp estimate-inaccuracy measures. It is in some ways more illuminating, at least for those familiar with Schervish (1989)'s construction of scoring rules. However it also has a high barrier to entry. Catrin Campbell-Moore and Daniel Rothschild later helped us figure out how to give a proof using only elementary methods (§B.3.1); we begin with this one.

Since Theorem 3.2 is local—i.e., concerns only a single random variable—we fix X for the rest of this appendix. We also adopt the following conventions:

- To save space, we set $\mathbb{E}_\pi(X) := e$.
- We let $\mathbb{E}(X)$ take values in $a_0 < \cdots < a_m$.
- We let X take values in $v_0 < \cdots < v_n$.

For convenience we restate the result here:

Theorem 3.2. *π totally trusts P with respect to X iff $\mathbb{E}_\pi(I_X(P)) \leq \mathbb{E}_\pi(I_X(e))$ for gsp scoring rules I_X, with equality if and only if $\pi(\mathbb{E}(X) = e) = 1$.*

Elementary-methods proof

For the elementary-methods proof, we first note that Campbell-Moore (2020) proves that every gsp estimate-inaccuracy measure is *value-directed*, which generalizes the constraint of truth-directedness to estimates generally: If $e_1 < e_2 \leq X(w)$ or $e_1 > e_2 \geq X(w)$, then $I_X(e_1, w) > I_X(e_2, w)$.

Next, we will need the following concept:

Monotone Strict Propriety Let I_X be a gsp. We say that I_X is *monotone strictly proper* if for any probability function π:

- If $e \leq s < t \leq v_n$, then $\mathbb{E}_\pi(I_X(t)) > \mathbb{E}_\pi(I_X(s))$
- If $v_0 \leq s < t \leq e$, then $\mathbb{E}_\pi(I_X(t)) < \mathbb{E}_\pi(I_X(s))$

The idea behind this definition is just that as estimates get closer to e, π expects them to be less inaccurate. Since for any gsp I_X, π expects its own estimate to be the least inaccurate, the following lemma is not surprising, but the proof (due to Catrin Campbell-Moore) is rather tricky.

Lemma B.15. *If I_X is a gsp, then I_X is monotone strictly proper.*

Proof. We show that if $e \leq s < t < v_n$, then $\mathbb{E}_\pi(I_X(t)) > \mathbb{E}_\pi(I_X(s))$. The second condition is entirely symmetric.

Fix s and t with $s < t$. We focus on probability functions ρ such that (i) if $v_i \leq s$, $\rho(X = v_i) \leq \pi(X = v_i)$, (ii) if $s < v_i < t$, $\rho(X = v_i) = \pi(X = v_i)$, and (iii) if $t \leq v_i$, $\rho(X = v_i) \geq \pi(X = v_i)$. Let Q be the set of all such ρ.

Since I_X is a gsp, it is value-directed. So, if $v_i \leq s$, $I_X(t, v_i) > I_X(s, v_i)$, and if $t \leq v_i$, $I_X(t, v_i) < I_X(s, V_i)$. So, for any $\rho \in Q$, $\mathbb{E}_\rho(I_X(t) - I_X(s)) \leq \mathbb{E}_\pi(I_X(t) - I_X(s))$.

We show that there exists a $\rho \in Q$ such that $\mathbb{E}_\rho(X) = s$. Let:

$$\rho^\star(X = v_i) = \begin{cases} 0 & \text{if } v_i \leq s \\ \pi(X = v_i) & \text{if } s < v_i < v_n \\ \pi(X = v_n) + \pi(X \leq s) & \text{if } v_i = v_n \end{cases}$$

Clearly $\rho^\star = \arg\max_{\rho \in Q} \mathbb{E}_\rho(X)$.

Given the definition of $\rho^\star$, we see $s < \mathbb{E}_{\rho^\star}(X)$. Note that $\pi \in Q$ and that Q is convex. So since $e \leq s \leq \mathbb{E}_{\rho^\star}(X)$, there is indeed some $\rho \in Q$ such that $\mathbb{E}_\rho(X) = s$. By strict propriety, $\mathbb{E}_\rho(I_X(t) - I_X(s)) > 0$, and as we've already established $\mathbb{E}_\pi(I_X(t) - I_X(s)) > \mathbb{E}_\rho(I_X(t) - I_X(s))$. This completes the proof. □

We can now prove Theorem 3.2 with elementary methods. The left-to-right direction is due to Daniel Rothschild.

Proof. We first prove the left-to-right direction. Given Total Trust, we show:

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) > e) < \mathbb{E}_\pi(I_X(e) \mid \mathbb{E}(X) > e) \tag{B1}$$

This suffices for the proof since a symmetric argument shows:

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) < e) < \mathbb{E}_\pi(I_X(e) \mid \mathbb{E}(X) < e) \tag{B2}$$

Jointly equations (B1) and (B2) entail this direction of the theorem.

To prove equation (B1): Let $w_1, \ldots, w_p$ be the worlds where for each i, $\mathbb{E}_i(X) > e$. Without loss of generality, assume that for each $i < p$, $\mathbb{E}_i(X) > \mathbb{E}_{i+1}(X)$. (In what follows, it will be clear that if $\mathbb{E}_w(X) = \mathbb{E}_{w'}(X)$ then they can be treated together.)

We will prove by induction for all k with $1 \leq k \leq p$ and for any $s < \mathbb{E}_k(X)$:

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) \geq \mathbb{E}_k(X)) < \mathbb{E}_\pi(I_X(s) \mid \mathbb{E}(X) \geq \mathbb{E}_k(X)) \tag{B3}$$

When $k = p$, equation (B3) entails (B1), since $\mathbb{E}_p(X)$ is the lowest value $\mathbb{E}(X)$ can take while still being greater than e.

Base case: $k = 1$, so $\mathbb{E}_k(X)$ is at its maximum value. Therefore,

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) \geq \mathbb{E}_1(X)) = \mathbb{E}_\pi(I_X(\mathcal{P}_1) \mid \mathbb{E}(X) \geq \mathbb{E}_1(X)) \tag{B4}$$

By total trust in X, we know that $\mathbb{E}_\pi(X \mid \mathbb{E}(X) \geq \mathbb{E}_1(X)) \geq \mathbb{E}_1(X)$. Since I_X is monotone strictly proper by Lemma B.15, we have then established the base case:

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) \geq \mathbb{E}_1(X)) < \mathbb{E}_\pi(I_X(s) \mid \mathbb{E}(X) \geq \mathbb{E}_1(X)) \tag{B5}$$

for any $s < \mathbb{E}_1(X)$.

Inductive Step: Suppose equation (B3) holds for all $k < i$. By Total Trust in X and monotone strict propriety, we have that for any $s < \mathbb{E}_i(X)$:

$$\mathbb{E}_\pi(I_X(\mathcal{P}_i) \mid \mathbb{E}(X) \geq \mathbb{E}_i(X)) < \mathbb{E}_\pi(I_X(s) \mid \mathbb{E}(X) \geq \mathbb{E}_i(X)) \tag{B6}$$

Since $\mathbb{E}_i(X) < \mathbb{E}_{i-1}(X)$, the inductive hypothesis tells us that:

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) \geq \mathbb{E}_{i-1}(X)) < \mathbb{E}_\pi(I_X(\mathcal{P}_i) \mid \mathbb{E}(X) \geq \mathbb{E}_{i-1}(X)) \tag{B7}$$

Since $[\mathbb{E}(X) \geq \mathbb{E}_i(X)] = [\mathbb{E}(X) = \mathbb{E}_i(X)] \cup [\mathbb{E}(X) \geq \mathbb{E}_{i-1}(X)]$, (B7) implies that

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) \geq \mathbb{E}_i(X)) < \mathbb{E}_\pi(I_X(\mathcal{P}_i) \mid \mathbb{E}(X) \geq \mathbb{E}_i(X)) \tag{B8}$$

Combining (B6) and (B8), we have that for any $s < \mathbb{E}_i(X)$:

$$\mathbb{E}_\pi(I_X(P) \mid \mathbb{E}(X) \geq \mathbb{E}_i(X)) < \mathbb{E}_\pi(I_X(s) \mid \mathbb{E}(X) \geq \mathbb{E}_i(X))$$

as desired.

To show the right-to-left direction: Suppose π does not totally trust P with respect to X. We consider the case where there exists t such that $\mathbb{E}_\pi(X \mid \mathbb{E}(X) \geq t) < t$ (as the case where there is a t such that $\mathbb{E}_\pi(X \mid \mathbb{E}(X) \leq t) > t$ is symmetric). Since $\mathbb{E}(X)$ can take only finitely many values $(a_0, \dots, a_m)$, there is some region (α, β) with $\alpha < \beta$ where Total Trust fails. I.e., for all $t \in (\alpha, \beta)$, $\mathbb{E}_\pi(X \mid \mathbb{E}(X) \geq t) < t$.

We construct a gsp where $\mathbb{E}_\pi(I_X(e)) < \mathbb{E}_\pi(I_X(P))$. For convenience (so that we only need six instead of nine cases in the scoring rule defined below), we choose α, β such that: there is no v_i, a_i in (α, β), and moreover $e \notin (\alpha, \beta)$. This is always possible since there are only finitely many v_i and a_i.

We define the following scoring rule for some (large) constant $C > 0$:

$$I_X(x, v_i) = \begin{cases} (x - v_i)^2 & \text{if } x \leq \alpha \text{ and } v_i < \alpha \\ (\alpha - v_i)^2 + C(x - \alpha)(x + \alpha - 2v_i) & \text{if } \alpha < x < \beta \text{ and } v_i < \alpha \\ (\alpha - v_i)^2 + C(\beta - \alpha)(\beta + \alpha - 2v_i) & \\ \qquad +(x - \beta)(x + \beta - 2v_i) & \text{if } x \geq \beta \text{ and } v_i < \alpha \\ (\alpha - x)(2v_i - \alpha - x) + C(\beta - \alpha)(2v_i - \beta - \alpha) & \\ \qquad +(\beta - v_i)^2 & \text{if } x \leq \alpha \text{ and } v_i > \beta \\ (\beta - v_i)^2 + C(\beta - x)(2v_i - \beta - x) & \text{if } \alpha < x < \beta \text{ and } v_i > \beta \\ (x - v_i)^2 & \text{if } x \geq \beta \text{ and } v_i > \beta \end{cases}$$

We leave it to the reader to verify that $I_X(x, v_i)$ is a gsp scoring rule.[67]

Suppose that $e < \alpha$. (The case where $e > \beta$ can be treated similarly.) Note that:

$$\mathbb{E}_\pi(I_X(P) - I_X(e)) = \sum_{i=0}^{n} \sum_{j=0}^{m} \pi(v_i, a_j)(I_X(a_j, v_i) - I_X(e, v_i)) \tag{B9}$$

where we write $\pi(v_i, a_j)$ for $\pi(X = v_i, \mathbb{E}(X) = a_j)$. We want to show that for C large enough (B9) is positive.

We can break (B9) up into four separate summations: (i) when $v_i, a_j < \alpha$, (ii) when $v_i < \alpha$, but $a_j > \beta$, (iii) when $v_i > \beta$, but $a_j < \alpha$, and (iv) when $v_i, a_j > \beta$. When we look at I_X, these correspond, respectively, to:

$$\sum_{v_i<\alpha} \sum_{a_j<\alpha} \pi(v_i, a_j)((a_j - v_i)^2 - (e - v_i)^2) \tag{B10}$$

$$\sum_{v_i<\alpha} \sum_{a_j>\beta} ((\alpha - v_i)^2 + (a_j - \beta)(a_j + \beta - 2v_i) + C(\beta - \alpha)(\beta + \alpha - 2v_i) - (e - v_i)^2)) \tag{B11}$$

$$\sum_{v_i>\beta} \sum_{a_j<\alpha} \pi(v_i, a_j)(a_j - e)(a_j + e - 2v_i)) \tag{B12}$$

$$\sum_{v_i>\beta} \sum_{a_j>\beta} \pi \ (v_i, a_j)((a_j - v_i)^2 - (a - e)(2v_i - \alpha - e) - (\beta - v_i)^2 + C(\beta - \alpha)(\beta + \alpha - 2v_i)) \tag{B13}$$

Summing these four expressions gives us (B9).

It's easy to see that we can ignore the contributions of (B10) and (B12), since C does not appear.

We also can ignore all terms in (B11) and (B13) that do not involve C. That is, we only need to consider:

$$\sum_{v_i<\alpha} \sum_{a_j>\beta} \pi(v_i, a_j)(C(\beta - \alpha)(\beta + \alpha - 2v_i)) + \sum_{v_i>\beta} \sum_{a_j>\beta} (C(\beta - \alpha)(\beta + \alpha - 2v_i)) \tag{B14}$$

For sufficiently large C, (B14) is positive if and only if (B9) is positive.

We divide (B14) by C and see that:

$$(\text{B14})/C = \sum_i \pi(v_i, \mathbb{E}(X) > \beta)((\beta - \alpha)(\beta + \alpha - 2v_i)) \tag{B15}$$

$$= \pi(\mathbb{E}(X) > \beta) \sum_i \pi(v_i \mid \mathbb{E}(X) > \beta)((\beta - \alpha)(\beta + \alpha - 2v_i)) \tag{B16}$$

$$= \pi(\mathbb{E}(X) > \beta)(\beta - \alpha)(\beta + \alpha - 2\mathbb{E}_\pi(X \mid \mathbb{E}(X) > \beta)) \tag{B17}$$

In (B17), $\pi(\mathbb{E}(X) > \beta) > 0$ and $\beta - \alpha > 0$. Given how we chose α, β, $\mathbb{E}_\pi(X \mid \mathbb{E}(X) > \beta) = \mathbb{E}_\pi(X \mid \mathbb{E}(X) > \alpha) < \alpha$. So since $\beta > \alpha$, $\beta + \alpha - 2\mathbb{E}_\pi(X \mid \mathbb{E}(X) > \beta) > 0$. So (B17) and therefore (B9) are both positive as desired. □

Schervish-style proof

We now turn to the slightly more in-depth but (we think) illuminating Schervish-style proof. For simplicity, we'll assume, without loss of generality, that perfect accuracy receives a score of 0. I.e., if X is a random variable, then $I_X(x, w) = 0$ if and only if $x = X(w)$, where $X(w)$ is X's value at w.

Schervish (1989) proves that for indicator variables, we can construct essentially arbitrary strictly proper scoring rules by placing various measures over the [0,1] interval as follows. (For an intuitive explanation of Schervish's theorem, see Levinstein 2017b.)

Theorem B.16 (Schervish 1989). *Let X be an indicator variable, and let $I_X(x, i)$ be a function from $[0, 1] \times \{0, 1\}$ to $\mathbb{R} \cup \{\infty\}$. Suppose $I_X(i, i) = 0$, and $I_X(x, i)$ is strictly increasing (decreasing) for $i = 0$ ($i = 1$), that I_X is continuous in its first argument over (0,1), and such that $I_X(i, j) = \lim_{t \to i} I_X(t, j)$ for $i, j = 0, 1$. Then I_X is a strictly proper scoring rule if and only if there exists a measure λ on [0,1] such that:*

$$I_X(x, 1) = \int_x^1 (1 - t)\, \lambda(\mathrm{d}t)$$

$$I_X(x, 0) = \int_0^x t\, \lambda(\mathrm{d}t)$$

for all x, where λ gives positive measure to every interval $[a, b)$ where $b > a$.

For example, if we let $\lambda(\mathrm{d}t) = 2\mathrm{d}t$, then $I_X(x, 1) = (1 - x)^2$, and $I_X(x, 0) = x^2$, which is the familiar Brier score.

Campbell-Moore (2020) generalizes Schervish's result to construct generalized strictly proper scoring rules for estimates.

Theorem B.17 (Campbell-Moore 2020). *Let X be a real-valued random variable such that $v_0 \leq X \leq v_n$, and let $I_X(x, k)$ be a function from $[v_0, v_n] \times [v_0, v_n]$ to $\mathbb{R}$. Suppose $I_X(x, x) = 0$ and $I_X(x, y)$ is strictly increasing as $|x - y|$ increases. Suppose further that $I_X(x, k)$ is absolutely continuous in its first argument over (v_0, v_n). Then I_X is a generalized strictly proper scoring rule iff there exists a measure λ on $[v_0, v_n]$ such that:*

$$I_X(x, k) = \int_k^x k - x\, \lambda(\mathrm{d}t)$$

for all x, where λ gives positive measure to every interval $[a, b)$ where $b > a$.

A few quick remarks. First, we require any gsp to be absolutely continuous in its first argument so that we can use the Lebesgue integral. (It is unclear if there is a way to relax this restriction.) Second, in the above result, we define the integral $\int_a^b f(t)\, \lambda(\mathrm{d}t) = -\int_b^a f(t)\, \lambda(\mathrm{d}t)$. This ensures $\int_k^x k - x\, \lambda(\mathrm{d}t) \geq 0$. Second, we've written the result so that X is bounded. This does not restrict us at all, since we're assuming all frames have finitely many worlds.

Campbell-Moore's result let's us easily generalize standard rules. For instance, if we again let $\lambda(\mathrm{d}t) = 2\mathrm{d}t$ over $[v_0, v_n]$ (where v_0 (v_n) is the minimum (maximum) value of X in the frame), then $I_X(x, k) = (k - x)^2$, which is the natural analog of the Brier score for estimates.

We now establish a useful lemma:

Lemma B.18. *If π totally trusts P with respect to X, then for all $t \in [0, 1]$:*

1. $\mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)\leq t}) - t \cdot \pi(\mathbb{E}(X) \leq t) \leq 0$;
2. $t \cdot \pi(\mathbb{E}(X) > t) - \mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)>t}) \leq 0$ *with strict inequality if* $\pi(\mathbb{E}(X) > t) > 0$.

Furthermore, if π does not totally trust P with respect to X, then there exist x, y with $x < y$ such that for all t in $[x, y]$ either:

3. $\mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)\leq t}) - t \cdot \pi(\mathbb{E}(X) \leq t) > 0$; *or*
4. $t \cdot \pi(\mathbb{E}(X) > t) - \mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)>t}) > 0$.

Proof. If $\pi(\mathbb{E}(X) \leq t) = 0$, then $\mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)\leq t}) = 0$, so (1) holds trivially. Likewise, if $\pi(\mathbb{E}(X) > t) = 0$, then $\mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)>t}) = 0$, so (2) holds trivially.

Suppose that $\pi(\mathbb{E}(X) \leq t) > 0$ and π totally trusts P. Then $\mathbb{E}_\pi(X \mid \mathbb{E}(X) \leq t) \leq t$. Expanding the definition of conditional expectation:

$$\sum_{i=0}^{n} \frac{\pi(X = v_i, \mathbb{E}(X) \leq t)}{\pi(\mathbb{E}(X) \leq t)} v_i \leq t \tag{B18}$$

Multiplying both sides by $\pi(\mathbb{E}(X) \leq t)$ and then appealing to the definition of expectation, we get that if π totally trusts P, then $\mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)\leq t}) - t \cdot \pi(\mathbb{E}(X) \leq t) \leq 0$.

We can obtain (2) through a similar derivation.

Now suppose π does not totally trust P. So, there exists some t such that $\mathbb{E}_\pi(X \mid \mathbb{E}(X) \leq t) > t$ or $\mathbb{E}_\pi(X \mid \mathbb{E}(X) > t) \leq t$. Suppose it's the former. Then for such a t:

$$\sum_{i=0}^{n} \frac{\pi(X = v_i, \mathbb{E}(X) \leq t)}{\pi(\mathbb{E}(X) \leq t)} v_i > t \tag{B19}$$

Again, multiplying both sides by $\pi(\mathbb{E}(X) \leq t)$ and appealing to the definition of expectation, we have $\mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)\leq t}) - t \cdot \pi(\mathbb{E}(X) \leq t) > 0$. Since $\mathbb{E}(X)$ can take on only finitely many values, $[\mathbb{E}(X) \leq t]$ is equivalent to either $[\mathbb{E}(X) \leq t + \varepsilon]$ or $[\mathbb{E}(X) \leq t - \varepsilon]$ for some sufficiently small ε. A parallel argument shows that if $\mathbb{E}_\pi(X \mid \mathbb{E}(X) > t) \leq t$, (4) holds. □

Lemma B.19. *If I_X is a generalized strictly proper scoring rule generated by measure λ and P is coherent, then:*

$$\mathbb{E}_\pi(I_X(P)) = \int_{v_0}^{e} \mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)\leq t}) - t \cdot \pi(\mathbb{E}(X) \leq t)\, \lambda(\mathrm{d}t) \tag{B20}$$

$$
\begin{aligned}
&+ \int_{e}^{v_n} t \cdot \pi(\mathbb{E}(X) > t) - \mathbb{E}_\pi(X \mathbb{1}_{\mathbb{E}(X)>t})\, \lambda(\mathrm{d}t) \\
&+ \mathbb{E}_\pi(I_X(e)) \\
= &\int_{e}^{v_0} t \cdot \pi(\mathbb{E}(X) > t) - \mathbb{E}_\pi(X \mathbb{1}_{\mathbb{E}(X)>t})\, \lambda(\mathrm{d}t) \qquad \text{(B21)} \\
&+ \int_{v_n}^{e} \mathbb{E}_\pi(X \mathbb{1}_{\mathbb{E}(X)\leq t}) - t \cdot \pi(\mathbb{E}(X) \leq t)\, \lambda(\mathrm{d}t) \\
&+ \mathbb{E}_\pi(I_X(e))
\end{aligned}
$$

Proof. By the definition of expectation and Theorem B.17, we have:

$$
\mathbb{E}_\pi(I_X(e)) = \sum_{i=0}^{n} \pi(X = v_i) \int_{e}^{v_i} (v_i - t)\, \lambda(\mathrm{d}t) \qquad \text{(B22)}
$$

(Recall: We are defining the integral so that $\int_e^{v_i} f(t)\,\lambda(\mathrm{d}t) = -\int_{v_i}^{e} f(t)\,\lambda(\mathrm{d}t)$.)

We now show that Equation (B20) holds. We have:

$$
\mathbb{E}_\pi(I_X(P)) = \sum_{i=0}^{n} \sum_{j=0}^{m} \pi(X = v_i, \mathbb{E}(X) = a_j) \int_{a_j}^{v_i} (v_i - t)\, \lambda(\mathrm{d}t) \qquad \text{(B23)}
$$

Fix v_i in the above equation and consider the inside summation:

$$
\sum_{j=0}^{m} \pi(X = v_i, \mathbb{E}(X) = a_j) \int_{a_j}^{v_i} (v_i - t)\, \lambda(\mathrm{d}t) \qquad \text{(B24)}
$$

First, divide up the a_j's so that $a_0 < \ldots < a_l \leq v_i$ and $v_i \leq a_{l+1} < \ldots < a_m$. So, we can re-write expression (B24) as:

$$
\sum_{j=0}^{l} \pi(X = v_i, \mathbb{E}(X) = a_j) \int_{a_j}^{v_i} (v_i - t)\, \lambda(\mathrm{d}t) \qquad \text{(B25)}
$$

$$
+ \sum_{j=l+1}^{m} \pi(X = v_i, \mathbb{E}(X) = a_j) \int_{a_j}^{v_i} (v_i - t)\, \lambda(\mathrm{d}t) \qquad \text{(B26)}
$$

Consider the first summand on line (B25). We first integrate from a_0 to v_i, then from a_1 (which is greater than a_0) to v_i, then from a_2 to v_i, etc. So we can re-write line (B25) as:

$$
\sum_{j=0}^{l} \pi(X = v_i, \mathbb{E}(X) \in \{a_0, \ldots, a_j\}) \int_{a_j}^{\min(a_{j+1}, v_i)} (v_i - t)\, \lambda(\mathrm{d}t)
$$

$$
\begin{aligned}
&= \sum_{j=0}^{l} \pi(X = v_i, \mathbb{E}(X) \leq a_j) \int_{a_j}^{\min(a_{j+1}, v_i)} (v_i - t)\, \lambda(\mathrm{d}t) \\
&= \sum_{j=0}^{l} \int_{a_j}^{\min(a_{j+1}, v_i)} \pi(X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\, \lambda(\mathrm{d}t) \\
&= \int_{a_0}^{v_i} \pi(X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\, \lambda(\mathrm{d}t) \\
&= \int_{v_0}^{v_i} \pi(X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\, \lambda(\mathrm{d}t) \qquad \text{(B27)}
\end{aligned}
$$

The second equality comes from the fact that if $a_j \leq t \leq a_{j+1}$, then $\pi(X = v_i, \mathbb{E}(X) \leq t) = \pi(X = v_i, \mathbb{E}(X) \leq a_j)$. The last equality comes from the fact that since P is coherent, $a_0 \geq v_0$. So, $\int_{v_0}^{a_0} \pi(X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\, \lambda(\mathrm{d}t) = 0$.

Applying a similar manipulation to line (B26), we have that the expression on line (B24) is equivalent to:

$$
\begin{aligned}
&\int_{v_0}^{v_i} \pi \;\; (X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\, \lambda(\mathrm{d}t) \\
&\qquad + \int_{v_i}^{v_m} \pi(X = v_i, \mathbb{E}(X) > t)(t - v_i)\, \lambda(\mathrm{d}t) \qquad \text{(B28)}
\end{aligned}
$$

Suppose first that $v_i < e$. Then we can re-write expression (B28) as:

$$
\begin{aligned}
&\int_{v_0}^{v_i} \pi \;\; (X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\, \lambda(\mathrm{d}t) \\
&\qquad + \int_{v_i}^{e} \pi(X = v_i, \mathbb{E}(X) > t)(t - v_i)\, \lambda(\mathrm{d}t) \\
&\qquad + \int_{e}^{v_m} \pi(X = v_i, \mathbb{E}(X) > t)(t - v_i)\, \lambda(\mathrm{d}t) \\
&= \int_{v_0}^{v_i} \pi \;\; (X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\, \lambda(\mathrm{d}t) \\
&\qquad + \int_{v_i}^{e} [\pi(X = v_i) - \pi(X = v_i, \mathbb{E}(X) \leq t)](t - v_i)\, \lambda(\mathrm{d}t) \\
&\qquad + \int_{e}^{v_m} \pi(X = v_i, \mathbb{E}(X) > t)(t - v_i)\, \lambda(\mathrm{d}t) \\
&= \int_{v_0}^{v_i} \pi \;\; (X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\, \lambda(\mathrm{d}t)
\end{aligned}
$$

$$
\begin{aligned}
&+\int_{v_i}^{e} \pi(X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\,\lambda(\mathrm{d}t) \\
&+\int_{e}^{v_i} \pi(X = v_i)(v_i - t)\,\lambda(\mathrm{d}t) \\
&+\int_{e}^{v_m} \pi(X = v_i, \mathbb{E}(X) > t)(t - v_i)\,\lambda(\mathrm{d}t) \\
= &\int_{v_0}^{e} \pi\ \ (X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\,\lambda(\mathrm{d}t) \qquad \text{(B29)} \\
&+\int_{e}^{v_m} \pi(X = v_i, \mathbb{E}(X) > t)(t - v_i)\,\lambda(\mathrm{d}t) \\
&+\int_{e}^{v_i} \pi(X = v_i)(v_i - t)\,\lambda(\mathrm{d}t)
\end{aligned}
$$

The second line on the right-hand side of the first equality comes from the law of total probability, since $\pi(X = v_i) = \pi(X = v_i, \mathbb{E}(X) \leq t) + \pi(X = v_i, \mathbb{E}(X) > t)$.

If $e < v_i$, we can show using a similar derivation that the expression (B28) is still equivalent to expression (B29). So, since expressions (B28) and (B29) are always equivalent, we can rewrite equation (B23) as:

$$\mathbb{E}_\pi(I_X(P)) = \sum_{i=0}^{n} \left(\int_{v_0}^{e} \pi(X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\,\lambda(\mathrm{d}t) \right. \qquad \text{(B30)}$$

$$+\int_{e}^{v_m} \pi(X = v_i, \mathbb{E}(X) > t)(t - v_i)\,\lambda(\mathrm{d}t) \qquad \text{(B31)}$$

$$\left. +\int_{e}^{v_i} \pi(X = v_i)(v_i - t)\,\lambda(\mathrm{d}t) \right) \qquad \text{(B32)}$$

First, note the summation of the integrals on line (B32) is equivalent to $\mathbb{E}_\pi(I_X(e))$, i.e.:

$$\sum_{i=0}^{n} \int_{e}^{v_i} \pi(X = v_i)(v_i - t)\,\lambda(\mathrm{d}t) = \mathbb{E}_\pi(I_X(e)) \qquad \text{(B33)}$$

Second, note that the first and second summations on lines (B30) and (B31) simplify as well to:

$$\sum_{i=0}^{n} \int_{v_0}^{e} \pi(X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\,\lambda(\mathrm{d}t) = \int_{v_0}^{e} \mathbb{E}_\pi(X \mathbb{1}_{\mathbb{E}(X) \leq t}) - t\pi(\mathbb{E}(X) \leq t)\,\lambda(\mathrm{d}t) \qquad \text{(B34)}$$

$$\sum_{i=0}^{n} \int_{e}^{v_m} \pi(X = v_i, \mathbb{E}(X) > t)(t - v_i)\,\lambda(\mathrm{d}t) = \int_{e}^{v_m} t\pi(\mathbb{E}(X) > t) - \mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)>t})\,\lambda(\mathrm{d}t) \quad \text{(B35)}$$

If we plug the right-hand sides of Equations (B33), (B34), and (B35) into lines (B30)–(B32), we see that Equation (B20) holds, which completes the proof of Equation (B20).

To see that Equation (B21) holds, note that in the derivation of line (B27) from line (B25), we integrated $v_i - t$ first from a_0 to v_i, then from a_1 to v_i, ..., and then from a_l to v_i. By the integration convention we've adopted, we instead could have integrated $t - v_i$ first from v_i to a_l, then from v_i to a_{l-1}, ..., then from v_i to a_0. Instead of line (B27), we would have:

$$\int_{v_i}^{v_0} \pi(X = v_i, \mathbb{E}(X) > t)(t - v_i)\,\lambda(\mathrm{d}t) \quad \text{(B36)}$$

And by an analogous treatment of line (B26), we would see that line (B24) is equivalent to:

$$\int_{v_i}^{v_0} \pi \;\; (X = v_i, \mathbb{E}(X) > t)(t - v_i)\,\lambda(\mathrm{d}t)$$
$$+ \int_{v_m}^{v_i} \pi(X = v_i, \mathbb{E}(X) \leq t)(v_i - t)\,\lambda(\mathrm{d}t)$$

The rest of the proof goes on to use a mirror image of the above derivation to obtain Equation (B21). □

Using these Lemmas, we're now in a position to prove our main accuracy result, Theorem 3.2.

Proof. We first prove the left-to-right direction. Suppose π totally trusts P. By Lemma B.19, Equation (B20) holds. By Fact (B.18), the first two terms in Equation (B20) are less than or equal to 0 and the last term is $\mathbb{E}_\pi(I_X(e))$. So $\mathbb{E}_\pi(I_X(P)) \leq \mathbb{E}_\pi(I_X(e))$ as desired.

To show that $\mathbb{E}_\pi(I_X(P)) = \mathbb{E}_\pi(I_X(e))$ when and only when $\pi(\mathbb{E}(X) = e) = 1$, consider the first two integrands in Equation (B20). As noted in Fact B.18, the second integrand will be less than 0 for all t such that $\pi(\mathbb{E}(X) > t) > 0$. It's also easy to see that the first integrand will be negative for all t such that both $\pi(\mathbb{E}(X) = t) = 0$ and $\pi(\mathbb{E}(X) \leq t) > 0$. Unless $\pi(\mathbb{E}(X) = e) = 1$, then, the integrands will be sometimes negative over some range (since there are only finitely possible values of $\mathbb{E}(X)$). Therefore, the inequality is strict when $\pi(\mathbb{E}(X) = e) < 1$.

To prove the right-to-left direction, suppose π does not totally trust P. We will show there is then some gsp scoring rule I_X such that $\mathbb{E}_\pi(I_X(P)) > \mathbb{E}_\pi(I_X(e))$.

Since π does not totally trust P, we appeal to items (3) and (4) of Fact B.18. There exists some interval $[x, y]$ such that at least one of the following holds:

1. $y \leq e$ and for all $t \in [x, y]$, $\mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)\leq t}) - t\pi(\mathbb{E}(X) \leq t) > 0$,
2. $y \leq e$ and for all $t \in [x, y]$, $t\pi(\mathbb{E}(X) > t) - \mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)>t}) > 0$
3. $e \leq x$ and for all $t \in [x, y]$, $\mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)\leq t}) - t\pi(\mathbb{E}(X) \leq t) > 0$, or
4. $e < x$ and for all $t \in [x, y]$, $t\pi(\mathbb{E}(X) > t) - \mathbb{E}_\pi(X\mathbb{1}_{\mathbb{E}(X)>t}) > 0$

We choose λ such that $\lambda([v_0, x])$ and $\lambda([y, v_m]) < \varepsilon$ with $\lambda([a, b]) = b - a$ for $a, b \in [x, y]$. If (1) or (4) hold, ε sufficiently small, we force the sum of the first two terms in Equation (B20) to be positive. If (2) or (3) hold, we force the sum of the terms in Equation (B21) to be positive. This completes the proof. □