The Humility Heuristic or: People Worth Trusting Admit to What They Don't Know

Mattias Skipper

Penultimate draft, forthcoming in *Social Epistemology*

Abstract: People don't always speak the truth. When they don't, we do better not to trust them. Unfortunately, that's often easier said than done. People don't usually wear a 'Not to be trusted!' badge on their sleeves, which lights up every time they depart from the truth. Given this, what can we do to figure out whom to trust, and whom not? My aim in this paper is to offer a partial answer to this question. I propose a heuristic—the "Humility Heuristic"—which is meant to help guide our search for trustworthy advisors. In slogan form, the heuristic says: *people worth trusting admit to what they don't know*. I give this heuristic a precise probabilistic interpretation, offer a simple argument for it, defend it against some potential worries, and demonstrate its practical worth by showing how it can help address some difficult challenges in the relationship between experts and laypeople.

Keywords: Epistemic humility, trust, testimony, expertise, epistemic trespassing

> *So I withdrew and thought to myself: "I am wiser than this man; it is likely that neither of us knows anything worthwhile, but he thinks he knows something when he does not, whereas when I do not know, neither do I think I know; so I am likely to be wiser than he to this small extent, that I do not think I know what I do not know."*
>
> — Socrates (Plato's *Apology*, 21d)

1. The Search for Trustworthy Advisors

One of the most salient facts about our epistemic lives is that we know much of what we know because others have told us. Most of us have never excavated any dinosaur fossils or detected any Higgs fields. Yet, many of us know that dinosaurs used to walk the earth and

that the Higgs field is all around us. We know this because others have done the requisite investigations and communicated their findings to us.

But despite the obvious benefits of knowledge sharing, the practice of relying on other people's say-so is fraught with pitfalls: lying (Fallis 2009), misleading (Stokke 2016), bullshitting (Frankfurt 2005 [1986]), and other forms of misinformation pervade social life.[1] Given that we live in a world of less than fully reliable advisors, each of us is confronted every day with a challenge of determining who deserves our trust. And it's a non-trivial challenge. People don't usually wear a 'Not to be trusted!' badge on their sleeves, which lights up every time they depart from the truth. The evidence we have to go on is much more scarce and indirect than that. Given this, what can we do to figure out whom to trust, and whom not?

My aim in this paper is to offer a partial answer to this question. I'll propose a heuristic (or "rule of thumb") which is meant to help guide our search for trustworthy advisors. In slogan form, the heuristic says:

The Humility Heuristic: People worth trusting admit to what they don't know.

I'll give this heuristic a precise probabilistic interpretation (§2), offer a simple argument for it (§3), defend it against some potential worries (§4), and demonstrate its practical worth by showing how it can help address some difficult challenges in the relationship between experts and laypeople (§5). The hope is that the considerations put forth will not only make it a little easier to separate the truthtellers from the bunch, but also serve to advance our understanding of the normative role of epistemic humility in our testimonial practices.

2. The Humility Heuristic in Probabilistic Terms

Our first task is to sharpen the heuristic. Consider an encounter between two agents: an "advisor" and an "advisee." The advisee is, we suppose, uncertain about whether a given proposition, p, is true. Fortunately (or not, as the case may be) the advisee is now given the opportunity to consult the advisor about whether, in his or her opinion, p is true.

[1] For a book length treatment of how misinformation can spread in societies, see O'Connor and Weatherall (2019). See also Hardwig (1985) and Lackey (2008) for some seminal entry points into the epistemological literature on testimony.

To analyze this situation in a precise manner, a bit of formal machinery will be helpful. Let P be the *rational credence function* of the advisee prior to consulting the advisor: that is, a function from propositions to numbers between 0% and 100%, representing the degrees of belief that the advisee *should* have at this initial point.[2] I'll make three assumptions about P.[3] First, I'll assume that P obeys the standard axioms of probability theory. Second, I'll assume that P obeys the Ratio Formula for conditional probabilities. Third, I'll assume that P is conditionalized on the advisee's background evidence (whatever it is). But apart from that, I won't make any controversial assumptions about what it takes for an agent's credences to be rational.

Next, we need to say something about what kinds of answers the advisor might give in response to the advisee's query. For the most part, I'll be focusing on two general kinds of answers that the advisor might give in response to a question of the form "Is p true?"

First, the advisor might answer "Yes." More generally, the advisor might testify to p by asserting that p is true. It won't matter for present purposes how, exactly, the assertion is made (whether it be made verbally, in writing, or through some other means of communication).[4] What matters is that the advisor outright asserts p in a way that is clear and unambiguous to the advisee. Henceforth, let's write "Tp" to denote the proposition "the advisor *T*estifies to p."

Second, the advisor might answer "I don't know." More generally, the advisor might admit to being *epistemically ignorant* about whether p is true. Again, the exact wording isn't important here (instead of saying "I don't know," the advisor might say "I couldn't tell you" or "I'll have to owe you an answer on that one").[5] Let's say that an agent who

---

[2] For the sake of simplicity, I'll assume that there is a *unique* rational credence function. While this is not in general an uncontroversial assumption, it should be harmless for present purposes. For further discussion of uniqueness, see White (2005), Schoenfield (2014), and Schultheis (2018).

[3] All three assumptions lie at the foundations of orthodox Bayesianism. See Bovens & Hartmann (2003) and Titelbaum (forthcoming) for some excellent background readings on Bayesian epistemology.

[4] For a detailed examination of what sets acts of assertion apart from other kinds of acts (and, in particular, other kinds of speech acts), see MacFarlane (2011).

[5] Note, in particular, that nothing turns on whether the advisor admits to lacking *knowledge* or whether she admits to lacking *justification to believe*. That is, rather than saying "I don't know," the advisor might as well say "I don't have sufficient evidence to answer that question." However, since it is much more common in ordinary discourse to talk about what we do or do not *know* than to talk about what we do or do not *have*

admits to not knowing whether a given proposition is true thereby expresses *epistemic humility* about that proposition, and let's write "*Hp*" to denote the proposition "the advisor expresses epistemic *H*umility about *p*."[6]

Of course, there are many other answers that an advisor might give in response to a question of the form "Is *p* true?" For example, rather than outright asserting *p*, the advisor might express a weaker kind of commitment to the truth of *p* by saying "I suspect that *p*" or "I'm fairly confident that *p*." As we'll see in §5, such "hedged" assertions raise interesting questions about the scope and limitations of the Humility Heuristic. But for now, I want to keep matters relatively simple by restricting attention to the answers described above.

With these preliminaries in place, we're ready for the official statement of the Humility Heuristic (where *p* and *q* are arbitrary propositions):[7]

Humility Heuristic: $P(p|Tp \,\&\, Hq) > P(p|Tp)$

(Slogan: people worth trusting admit to what they don't know.)

The Humility Heuristic says that the advisee should treat *Tp* & *Hq* as stronger evidence for *p* than *Tp* alone. More precisely, it says that the advisee's credence in *p* given that the advisor testifies to *p and* admits to not knowing whether *q* is true should be higher than the advisee's credence in *p* given that the advisor testifies to *p*.

---

justification to believe, I'll stick to the locution "I don't know" as the paradigmatic way of expressing the kind of epistemic humility that I'm interested in.

[6] A remark on terminology here: the term "epistemic humility" (together with its close cousin "epistemic modesty") has been given a number of different meanings in the philosophical literature. For example, Elga (2016) stipulates that you're "epistemically humble" iff you're uncertain about whether your beliefs will converge to the truth given enough evidence, and Dorst (2019) stipulates that you're "epistemically modest" iff you're uncertain about what it is rational for you to believe. My usage of the term "epistemic humility" differs from both Elga's and Dorst's. On my usage, you express epistemic humility about *p* iff you admit to not knowing *p*. Note, however, that all three notions are used as (semi-)technical terms, not competing analyses of the same intuitive concept. In particular, my usage of the word "humility" isn't supposed to track our ordinary intuitions about humility as a virtue that admits of excess as well as deficiency. As an anonymous referee rightly points out, there is an intuitive sense in which someone who says "I don't know" in response to every question isn't humble, but intellectually timid. For present purposes, however, I'll stipulate that such a person would indeed express a high degree of epistemic humility.

[7] Here is an equivalent formulation of the Humility Heuristic, which some readers may find easier to parse: $P(p|Tp \,\&\, Hq) > P(p|Tp \,\&\, \sim Hq)$.

Let me clarify a few points about the Humility Heuristic. First, note that the Humility Heuristic is a purely ordinal claim: it says *that* $P(p|Tp \,\&\, Hq)$ is greater than $P(p|Tp)$, but it says nothing about *how much* greater $P(p|Tp \,\&\, Hq)$ is than $P(p|Tp)$. In other words, all the Humility Heuristic says is that people who admit to what they don't know are at least slightly more trustworthy for that reason. Of course, it's natural to wonder whether and how the heuristic may be strengthened. I'll briefly touch on this question in §4. However, a detailed investigation must wait for another occasion. The aim of this paper is just to establish the purely ordinal claim. As I hope to be able to demonstrate, this would be a significant step forward in its own right.

Second, note that there are various probability claims in the vicinity of the Humility Heuristic that might be thought to follow from the heuristic, but which *don't*. Here are two examples:

(a) $P(p|Tp) > P(p)$

(b) $P(p|Hq) > P(p)$

Neither (a) nor (b) follows from the Humility Heuristic. In fact, it is possible for the Humility Heuristic to be accurate even if neither Tp nor Hq supports p.[8] For purposes of illustration, however, I'll focus mainly on cases where Tp provides at least *some* evidence for p, in which case the Humility Heuristic implies that $Tp \,\&\, Hq$ provides *even stronger* evidence for p.

Third, note that the Humility Heuristic *iterates*: that is, the advisee should become (at least slightly) more trusting in the advisor each time the advisee learns that the advisor has expressed humility about some proposition. This is due to the fact that P is conditionalized on the advisee's background evidence, which may include evidence about the advisor having expressed humility on previous occasions. To illustrate, let q and r be two propositions that the advisor currently hasn't expressed humility about, and let P be the advisee's rational credence function at this stage. We can then imagine that the advisee undergoes a series of learning experiences. First, the advisee learns Hq and updates her credence in p to $P_{Hq}(p) = P(p|Hq)$. Then the advisee learns Tp and updates her credence to

[8] Here is a quick proof: we define a probability distribution over the set of propositional variables $\{p, Tp, Hq\}$ such that $P(p) = .5$, $P(Hq) = .4$, $P(Tp) = .2$, $P(p|Hq) = P(p|Tp) = .5$, $P(Tp \,\&\, Hq) = .1$, and $P(p|Tp \,\&\, Hq) = 1$. Given this, $P(p|Tp \,\&\, Hq) > P(p|Tp)$, $P(p|Hq) = P(p)$, and $P(p|Tp) = P(p)$, which means that the Humility Heuristic is accurate, although neither (a) nor (b) obtains.

$P_{Hq}(p|Tp)$. Finally, she learns Hr and updates her credence to $P_{Hq}(p|Tp \,\&\, Hr)$. Given the Humility Heuristic, it follows that $P_{Hq}(p|Tp \,\&\, Hr) > P_{Hq}(p|Tp) > P(p|Tp)$. Thus, since $P_{Hq}(\cdot) = P(\cdot|Hq)$, we get that $P(p|Tp \,\&\, Hq \,\&\, Hr) > P(p|Tp \,\&\, Hq) > P(p|Tp)$.

Finally, keep in mind that the Humility Heuristic is intended as a *heuristic*. There is nothing probabilistically incoherent about a credence function that violates the inequality $P(p|Tp \,\&\, Hq) > P(p|Tp)$, for some p and q.[9] The question we'll be interested in is whether the Humility Heuristic is *typically* accurate in the kinds of epistemic situations that we may realistically find ourselves in. As I'll argue in the next section, I think this question can be given a positive answer.

But why care to provide an argument for the Humility Heuristic in the first place? I suspect that many readers will find the Humility Heuristic intuitively plausible (as I myself do). So, in defending the Humility Heuristic, I don't take myself to take a stance on a controversial issue. Nevertheless, I believe that there is something valuable to be gained from providing a careful philosophical analysis of the Humility Heuristic. It can often be interesting and illuminating to search for a theoretical vindication of a claim, even if that claim is presumed to be true at the outset. That's the spirit in which the ensuing discussion is to be taken.[10]

3. An Argument for the Humility Heuristic

The backbone of the argument is the following result:

> **Sufficiency Result:** The Humility Heuristic is accurate if the following three conditions obtain:
>
> **C1** $P(Tp|{\sim}p \,\&\, Hq) < P(Tp|{\sim}p)$

[9] The easiest way to see this is to let the unconditional probability of p be extreme: that is, to assume that $P(p) = 1$ or $P(p) = 0$. In either case, it follows that $P(p|Tp \,\&\, Hq) = P(p|Tp)$, since extreme probabilities are preserved conditional on any new evidence.

[10] As an anonymous referee has rightly pointed out to me, it's natural to think that the Bayesian approach taken in this paper may be complemented by resources from the literature on virtue epistemology. I very much welcome attempts at exploring the Humility Heuristic from a virtue epistemological perspective, but doing so is beyond the scope of this paper. Readers who are interested in pursuing this line of investigation may want to consult, e.g., Battaly (2008), Cassam (2016), and Whitcomb et al. (2015).

(The advisee should consider it more likely that the advisor testifies to p given that $\sim p$ than given that $\sim p$ *and* the advisor admits to not knowing whether q.)

C2 $P(p|Hq) \geq P(p)$
(The advisee's credence in p given that the advisor admits to not knowing whether q shouldn't be lower than the advisee's unconditional credence in p.)

C3 $P(Tp|Hq) \geq P(Tp)$
(The advisee's credence that the advisor will testify to p given that the advisor admits to not knowing whether q shouldn't be lower than the advisee's unconditional credence that the advisor will testify to p.)

This result is simply a theorem of the probability calculus (a proof is included in the Appendix). Nevertheless, it holds valuable information about the conditions under which the Humility Heuristic is accurate: it tells us that the Humility Heuristic is accurate whenever a certain set of conditions obtain. The question, then, is when these conditions obtain. Below I go over each of the conditions, explaining what they say, what role they play in establishing the Sufficiency Result, and why we should expect them to obtain in most (although not all) ordinary situations.

As we'll see, there are some worries one might have about each of the conditions as well as about the Humility Heuristic itself. I'll address some of these worries as we go along, but I'll defer the worries that I take to run a bit deeper until §4, when the positive case for the Humility Heuristic is on the table.

3.1. Condition 1: $P(Tp|\sim p \,\&\, Hq) < P(Tp|\sim p)$

The first condition is also the most critical one, for reasons that will become clear. It says, roughly, that people who are willing to admit to what they don't know are less likely to make false assertions than people who are *not* willing to admit to what they don't know. More precisely, it says that the advisee should consider it more likely that the advisor testifies to p given that p is false than given that p is false *and* the advisor admits to not knowing whether q is true.

The rationale behind this condition is fairly straightforward: presumably, someone who is willing to admit to not knowing whether a given proposition is true will also be more likely, other things being equal, to admit to not knowing *various other* unknown propositions—compared, that is, to someone who *isn't* willing to admit to not knowing

whether said proposition is true. After all, the fact that someone admits to not knowing whether a given proposition is true is typically at least a weak indication of a general aversion against making false assertions. So, the fact that the advisor expresses epistemic humility about *q* is typically going to be at least a weak *pro tanto* reason for the advisee to think that the advisor wouldn't assert *p*, if *p* were false.[11]

To illustrate the point, consider the following example:

> **Press Conference:** You're at a press conference in the Ministry of Foreign Affairs, sitting alongside the rest of the press corps. When called upon, you're allowed to ask two questions directed to the foreign minister. You've decided to ask the following two questions:
>
> **Q1** "Does Country X possess weapons of mass destruction?"
>
> **Q2** "Would policy Y, if implemented, have effect Z?"
>
> In response, the foreign minister provides the following answers:
>
> **A1** "I'm afraid we don't know enough to answer that question."
>
> **A2** "Yes, it would."

We can then ask: how should you take the fact that the foreign minister expresses epistemic humility about the subject-matter of Q1 to bear on whether her answer to Q2 is correct? The answer to this question clearly depends on your background evidence. But on most realistic ways of filling in the details of the case, you should presumably treat the fact that the foreign minister is willing to admit to not knowing the answer to Q1 as at least a weak *pro tanto* reason to think that she wouldn't have answered "Yes" in response to Q2, if the true answer had been "No." After all, the fact that the foreign minister is willing to express epistemic humility about the subject-matter of Q1 makes it at least slightly less likely that she is systematically lying or bullshitting or otherwise being insensitive to the truth on this occasion.[12] And that's all it takes for C1 to obtain.

---

[11] Doesn't this depend on the content of *p* and *q*? In particular, doesn't it depend on whether *p* and *q* fall within the same general domain? The short answer is "No." I'll return to the issue in §4.1.

[12] Of course, the foreign minister might be lying about whether she knows the answer to the first question. But that's a subtly different matter. It's one thing to lie about *p*; it's another thing to lie about whether you know *p*. Someone who lies about not knowing *p* doesn't thereby make a false assertion about *p*. As such, it's not clear that the possibility that the foreign minister lies about not knowing the answer to the first question has any significant bearing on the probability that her answer to the second question is false. But in any case, I doubt that this possibility will create problems for C1 in most ordinary situations.

I submit that most ordinary situations are like Press Conference in this respect. That is to say, it is typically reasonable to treat the fact that a person expresses epistemic humility about a given proposition as at least a weak indication of a general aversion against making false assertions.

I say "typically" because there may be exceptions. Suppose, for example, that you have good reason to think that it would be in your friend's interest to lie about who invented the light bulb, but not in your friend's interest to lie about who is the current president of Switzerland (perhaps because you have good reason to think that your friend, being an aficionado of 19th century technology, would be embarrassed by not knowing who invented the light bulb, but not embarrassed by not knowing who is the current president of Switzerland). If that's your situation, the fact that your friend admits to not knowing who is the current president of Switzerland might not give you any reason to think that your friend won't lie about who invented the light bulb. Or suppose you have good reason to think that your friend is subject to what we might call "forced admission of ignorance:" situations in which there's no option but to admit one's ignorance about some matter. For example, we can imagine that your classmate is asked by your French teacher what "L'éducation est un droit de l'homme" means. If your friend doesn't know the answer and sees himself forced to admit as much, this presumably doesn't give you any reason to think that your friend won't lie in situations where this option is available.[13]

But even if C1 isn't immune to counterexamples, it can still do its job in establishing the Humility Heuristic as a good rule of thumb. What matters for this purpose is that C1 *typically* obtains—and that's what I take to be plausible on the grounds that it typically seems reasonable to treat the fact that someone is willing to admit to what they don't know as at least a weak indication of a general aversion against making false assertions.

3.2. Condition 2: $P(p|Hq) \geq P(p)$

The second condition plays a somewhat more peripheral role. It says, roughly, that the fact that the advisor admits to not knowing whether q is true doesn't constitute direct evidence against p. More precisely, it says that the advisee's credence in p given that the advisor admits to not knowing whether q is true shouldn't be lower than the advisee's unconditional credence in p.

[13] Thanks to an anonymous referee for drawing my attention to this kind of phenomenon.

The reason why C2 is needed to establish the Humility Heuristic is that, if *Hq* constitutes direct evidence against *p*, then *Tp* & *Hq* can fail to be stronger evidence for *p* than *Tp* alone, simply because *Hq* acts as a rebutting defeater of *p*. Suppose, for example, that you have good reason to think that your friend would have known *q*, if *p* had been true (perhaps because you have good reason to think that someone would have told your friend that *q*, if *p* had been true).[14] If that's your situation, you should take the fact that your friend admits to not knowing whether *q* is true to constitute evidence against *p*. After all, if *p* had been true, your friend would most likely have known *q*, in which case he would most likely not have admitted to not knowing whether *q* is true. So, assuming that *Hq* is a strong enough rebutter of *p*, this is a case where *Tp* & *Hq* doesn't support *p* more strongly than *Tp* alone.

But again, what matters for present purposes is whether C2 *typically* obtains. And I think it does. Perhaps the easiest way to see this is by noticing that C2 will, at the very least, obtain whenever *Hq* is evidentially irrelevant to *p*: that is, when *Hq* neither raises nor lowers the probability of *p* relative to the advisee's background evidence. This already covers a wide range of ordinary cases. To mention just a few mundane examples: the fact that your colleague admits to not knowing whether the Lakers beat the Celtics last night seems to have no evidential bearing on whether Paris is the capital of France; the fact that your teacher admits to not knowing who was awarded the inaugural Fields Medal seems to have no evidential bearing on whether the chemical structure of water is H_2O; and so on. More generally: unless the advisee has special reason to think that the question of whether the advisor knows *q* has a direct evidential bearing on whether *p* is true, C2 will (*a fortiori*) obtain.

3.3. Condition 3: $P(Tp|Hq) \geq P(Tp)$

The third condition also plays more of a peripheral role. It says, roughly, that the fact that the advisor admits to not knowing whether *q* is true doesn't make it any less likely that the advisor will testify to *p*. More precisely, it says that the advisee's credence that the advisor will testify to *p* given that the advisor admits to not knowing whether *q* is true shouldn't be lower than the advisee's unconditional credence that the advisor will testify to *p*.

[14] This example is inspired by Goldberg's (2010, ch. 6) discussion of inferences from "absence of evidence" to "evidence of absence."

The reason why C3 is needed to establish the Humility Heuristic is a little more subtle: in some cases, if *Hq* is evidence against *Tp*, the Humility Heuristic will fail to be accurate, even if C1 and C2 both obtain. Suppose, for example, that you're about to ask your friend two questions: (i) "What is the capital of France?," and (ii) "What is the capital of Italy?" Suppose also that, given your background evidence about people's general knowledge about European geography, you find it highly unlikely that your friend would know the capital of France, but fail to know the capital of Italy. If that's your situation, your credence that your friend will assert that Paris is the capital of France given that your friend admits to not knowing the capital of Italy should be lower than your unconditional credence that your friend will assert that Paris is the capital of France. After all, the fact that your friend doesn't know the capital of Italy is a strong indication (relative to your background evidence) that your friend doesn't know the capital of France either.

Now, let's ask: should you, as the Humility Heuristic dictates, be less confident that Paris is the capital of France given that your friend asserts that Paris is the capital of France than given that your friend asserts that Paris is the capital of France *and* admits to not knowing the capital of Italy? Presumably not. After all, you should find it highly unlikely in advance that your friend would know the capital of France, but fail to know the capital of Italy. Thus, you should take the fact that your friend both asserts that Paris is the capital of France and admits to not knowing the capital of Italy to be a strong indication that your friend is either confused or insincere or otherwise insensitive to the truth on this occasion. So, this is a case where C3 fails to obtain, and where, as a consequence, the Humility Heuristic fails to be accurate.

Once again, however, there are general grounds for thinking that C3 *typically* obtains. The reasoning is similar to that offered in favor of C2: C3 will, at the very least, obtain whenever *Hq* is evidentially irrelevant to *Tp* relative to the advisee's background evidence. And this covers a wide range of ordinary cases: the fact that your mother admits to not knowing who founded Marlboro seems to have no evidential bearing on the question of whether she will tell you that it will be rainy tomorrow; the fact that your business partner admits to not knowing who arranged last year's office party seems to have no evidential bearing on whether she will tell you that today's meeting is cancelled; and so on. More generally: unless the advisee has special reason to think that the question of whether the

advisor knows *q* has a direct evidential bearing on whether the advisor will testify to *p*, C3 will (*a fortiori*) obtain.[15]

3.4. Beyond the Sufficiency Result

We've now seen that the Humility Heuristic is accurate whenever C1-C3 obtain. But what happens when they *don't*? Is the Humility Heuristic inaccurate in all such cases? No. Just as none of the three conditions is individually *sufficient* for the Humility Heuristic to be accurate, none of them is individually *necessary* either. In fact, the strongest logical combination of C1-C3 that is necessary for the Humility Heuristic to be accurate is their disjunction:

> **Necessity Result:** The Humility Heuristic is accurate only if at least one of C1-C3 obtains.

Like the Sufficiency Result, the Necessity Result is a theorem of the probability calculus.[16] It tells us that the Humility Heuristic is guaranteed to be inaccurate if C1-C3 all fail to obtain at the same time. Now, if the foregoing remarks are basically correct, we should expect this rarely to be the case. But there is another result in the vicinity, which promises wider applicability:

> **Equivalence Result:** The Humility Heuristic is equivalent to C1 if the following conditions obtain:

---

[15] Here is a slight complication: I've said that the fact that an advisor expresses epistemic humility on a given occasion is typically at least a weak indication of a general aversion against making *false* assertions. By the same token, doesn't the fact that an advisor expresses epistemic humility on a given occasion typically provide at least a weak indication of a general aversion against making assertions *simpliciter*? And doesn't this in turn generate a broad class of counterexamples to C3? That may well be so. However, the relevant class of counterexamples to C3 won't carry over as counterexamples to the Humility Heuristic. When a counterexample to C3 constitutes a counterexample to the Humility Heuristic, it is because it describes a situation in which the fact that the advisor expresses epistemic humility about *q* makes it more likely that the advisor would falsely assert *p*, if she were to assert *p* at all. That's what made the "European geography" case discussed above a counterexample to the Humility Heuristic. But the counterexamples to C3 under consideration here don't share this feature with the European geography case. They simply describe cases in which the fact that the advisor expresses epistemic humility about *q* makes it less likely that the advisor will assert *p* in the first place.

[16] The proofs of this result and the next are similar to the proof of the Sufficiency Result in the Appendix. The details are left out here.

C2* $P(p|Hq) = P(p)$

C3* $P(Tp|Hq) = P(Tp)$

This result tells us that C1 is both necessary and sufficient for the Humility Heuristic to be accurate provided that we replace C2 and C3 by two stronger conditions, C2* and C3*, which say that Hq is evidentially irrelevant to both p and Tp. In a trivial sense, since C2* and C3* are logically stronger than C2 and C3, they will obtain less often. But we should nevertheless expect them to obtain in a fairly wide range of ordinary situations, for much the same reason that we should expect C2 and C3 to obtain in a wide range of ordinary situations: it's often reasonable to assume that Hq has no direct evidential bearing on p and Tp. Whenever this is the case, the Equivalence Result tells us that the question of whether the Humility Heuristic is accurate comes down to whether C1 obtains. That's why I said earlier that C1 can be seen as the most critical of the three conditions.

4. Worries about the Humility Heuristic

I find the case in favor of the Humility Heuristic compelling. Nevertheless, there are some worries one might have about it. In this section, I'll look at three of the most interesting worries that have come to my attention. Ultimately, I don't think either worry has much force against the heuristic, but they each raise important questions about its scope and limitations worth examining in their own right.

4.1. Domain-Relative Trustworthiness

The first worry goes as follows:

> The Humility Heuristic, as stated, doesn't say anything about whether p and q must fall within the same general domain. Yet, people's degree of trustworthiness clearly varies from domain to domain: someone who is trustworthy on matters of cosmology need not be trustworthy on matters of developmental psychology; someone who is trustworthy on matters of English literature need not be trustworthy on matters of US foreign politics; and so on. In general, someone who is trustworthy in one domain need not be trustworthy in other, far removed domains. Doesn't this mean that we should expect the Humility Heuristic to be accurate only when p and q fall within the same domain, or at least suitably similar domains?

There is clearly something right about the initial observation that people's degree of trustworthiness varies from domain to domain. Of course, one can quibble about how to individuate domains. But that's beside the point here. Regardless of how we choose to individuate domains, people's degree of trustworthiness is presumably going to vary from domain to domain. The question is whether this elementary fact spells trouble for the Humility Heuristic. That's where I think the worry misfires.

The thing to keep in mind here is that the Humility Heuristic is a purely ordinal claim: it says *that* Tp & Hq supports p more strongly than Tp alone, but it doesn't say anything about *how much* more strongly Tp & Hq supports p than Tp alone. The relevant question, then, is whether this purely ordinal claim is (typically) true in cases where p and q fall within very different domains. And I think this question can be given a positive answer.

The easiest way to see this is by looking at the main condition, C1, which says that the advisor is less likely to assert p given $\sim p$ & Hq than given $\sim p$ alone. Should we expect this condition to be satisfied in cases where p and q fall within very different domains? More specifically, should we expect it to be satisfied if the advisor is much less trustworthy relative to the p-domain than the q-domain?

Insofar as the foregoing remarks on C1 are correct, the answer "Yes." Even if p and q fall within very different domains, the fact that the advisor admits to not knowing whether q is true is still at least a weak indication of a general aversion against making false assertions, including about matters within the p-domain. Suppose, for example, that you (as seems reasonable) consider your physics professor to be more trustworthy on matters of cosmology than on matters of developmental psychology. Suppose also that, on a given occasion, your physics professor admits to not knowing whether the universe has a flat or curved geometry. Should this expression of epistemic humility about the geometry of the universe make you more confident that your professor won't make false assertions about matters related to developmental psychology? Presumably so. For you should take the fact that your professor is willing to admit to not knowing whether the universe has a flat or curved geometry to be at least a weak indication of a general aversion against making false assertions, including about matters related to developmental psychology. This is not to say that your professor's expression of epistemic humility about the geometry of the universe raises her degree of trustworthiness on matters of developmental psychology by a *large amount* (indeed, that may seem doubtful). The claim is just that your professor's expression of epistemic humility about the geometry of the universe makes it at least

slightly less likely that they will make false assertions about matters of developmental psychology.

Nevertheless, I think the present worry brings out an interesting point about when the Humility Heuristic may prove most *useful*. Suppose we wanted to go beyond the purely ordinal claim and say something about when the difference between $P(p|Tp \,\&\, Hq)$ and $P(p|Tp)$ is most pronounced. If that were our goal, we'd do well to pay attention to the specific content of p and q, and, in particular, whether p and q fall within suitably similar domains. But the fact that people's degree of trustworthiness tends to vary from domain to domain doesn't cause trouble for the Humility Heuristic, understood as a purely ordinal claim.

4.2. Hedged Assertions

The second worry I'd like to consider goes as follows:

> The Humility Heuristic doesn't specify whether p and q must be *distinct* propositions. But the heuristic doesn't seem to provide accurate guidance in cases where p and q are *identical*. The problem is not so much to do with "Moorean" assertions of the form "p, but I don't know p." Such assertions are arguably quite rare anyway. Rather, the trouble is to do with "hedged" assertions such as "I believe she's gonna make it, but I might be wrong" or "I suspect he committed the crime, but I don't know for sure." Such assertions are pervasive in ordinary discourse, and their logical form seems to be well captured by the conjunction "$Tp \,\&\, Hp$." However, hedged assertions, by their nature, serve to express a relatively *weak* kind of commitment to the truth of the asserted proposition, thereby providing the hearer with a correspondingly weak reason to believe the asserted proposition. Doesn't this place a rather significant limitation on the scope of the Humility Heuristic?

I think this worry is basically sound, except for one key point: the logical form of a hedged assertion is not well captured by the conjunction "$Tp \,\&\, Hp$." Recall that the intended interpretation of "Tp" is as an *outright* assertion of p. And a hedged assertion like "I believe she's gonna make it, but I might be wrong" presumably doesn't contain an outright assertion in its first conjunct (despite surface appearances to the contrary). As such, the Humility Heuristic was never supposed to say anything about hedged assertions

in the first place. In particular, it shouldn't be taken to imply that a hedged assertion such as "I believe she's gonna make it, but I might be wrong" gives you more of a reason to believe that she's gonna make it than the corresponding outright assertion "I believe she's gonna make it."

That said, I find the present worry illuminating, because it reminds us that we need to be careful about how we generalize the Humility Heuristic, should we want to do so. In particular, it shows that we can't straightforwardly generalize the Humility Heuristic to cover hedged assertions without running into a broad class of counterexamples.

4.3. Exploiting the Humility Heuristic

The third worry goes as follows:

> When an advisor knows (or has reason to suspect) that an advisee adheres to the Humility Heuristic, the advisor can exploit this fact by expressing humility about one matter in order to gain the advisee's trust on another matter of interest. For example, a corrupt politician who is interested in helping the oil industry might express humility about, say, the effects of free education on public health in order to gain the public's trust on matters having to do with the effects of fossil fuel combustion on the global climate. Doesn't this possibility of exploitation count against implementing the Humility Heuristic?

I think the right thing to say here is that the possibility of exploitation does indeed count against implementing the Humility Heuristic, but that it doesn't outweigh the potential benefits of the heuristic. Generally speaking, we can distinguish between two types of exploitation scenarios here.

In the first type of scenario, the advisee knows (or has reason to suspect) that the advisor engages in the relevant kind of exploitation. In such scenarios, C1 fails to obtain: the advisee shouldn't take the advisor's expression of humility to be indicative of a general aversion against making false testimony. As such, this type of exploitation scenario can be added to the set of counterexamples to C1 already mentioned in §3.1.

In the second type of scenario, the advisee doesn't know (or have reason to suspect) that the advisor engages in exploitation. This type of scenario is perhaps more disturbing than the first, since there is no obvious way for the advisee to guard him- or herself against such exploitation. However, it seems to me that it would be an overreaction to refrain from

implementing the Humility Heuristic in order to avoid this problem. After all, many other valuable heuristics can be exploited as well. For instance, to take an example that will connect up with the next section, it seems perfectly rational for a layperson to take the fact that a putative expert in a given field has been appraised by other experts in the field as a *pro tanto* reason to think that the putative expert is, indeed, a genuine expert (cf. Goldman 2001). Yet, this heuristic can be exploited: a corrupt person may bribe a group of experts to provide the relevant kind of appraisal, thereby deceiving the advisee(s). This risk of exploitation undeniably counts against implementing this "expert appraisal" heuristic to some extent. But it hardly outweighs the potential benefits of the heuristic. The same, I say, goes for the Humility Heuristic.

5. Putting the Humility Heuristic to Work: Experts vs. Laypeople

In the hope of demonstrating the practical significance of what has been said so far, I'd like to close by looking in more depth at a specific application of the Humility Heuristic. I've chosen to focus on a set of issues that arise in the relationship between experts and laypeople. There are, no doubt, many other applications that deserve a separate discussion. But I hope that the discussion below will touch on challenges that many readers will be able to recognize from their own epistemic lives.

We begin with a bit of stage setting: it's well-known that expert testimony plays a central role in communities with a high degree of division of cognitive labor.[17] Yet, the dissemination of knowledge by expert testimony is complicated by the fact that experts don't always agree amongst themselves. When they don't, it can be difficult for the laypeople among us to figure out who is in the right. After all, we're usually not in a position to adjudicate expert disagreements by looking at the relevant first-order evidence and arguments ourselves. We simply don't have the requisite knowledge and competencies to do so.

In his seminal discussion of this "novice/expert" problem, Goldman (2001, p. 94) introduces a helpful distinction between *esoteric* and *exoteric* information in an expert's discourse. Esoteric information belongs to the relevant area of expertise, and hence isn't the kind of information that laypeople are usually in a good position to rely on. Exoteric information, on the other hand, doesn't belong to the relevant area of expertise, and hence

[17] See Hardwig (1985) and Kitcher (1990; 1993) for some excellent discussions on this point.

is more readily accessible to the layperson. As Goldman himself points out, this distinction clearly isn't sharp—it admits of degrees. But for present purposes, it won't hurt to talk about esoteric and exoteric information in binary terms.

Now, the central lesson of Goldman's discussion is that even if laypeople can't rely on esoteric information to adjudicate expert disagreements they might nevertheless rely on various kinds of exoteric information to make an informed judgment about which expert is most worthy of being trusted. Goldman himself discusses five broad categories of exoteric information (including "expert appraisal," as mentioned in the previous section), which I won't rehearse here. Instead, I'd like to draw attention to a different kind of exoteric information, which has been brought out by Dellsén (2016).

Dellsén argues that the fact that there is disagreement amongst a group of experts on a given issue is a *pro tanto* reason for laypeople to trust the experts on issues on which they agree. As he puts it: "expert disagreement supports the consensus." For example, if I learn that a group of cosmologists disagree amongst themselves about whether the universe has a flat or curved geometry, I should, Dellsén submits, treat this fact as a *pro tanto* reason to trust their consensus (assuming that there is one) on the age of the universe. Needless to say, one might take issue with this claim. But if it is right, it says something interesting and important about expert disagreement, namely that it itself can be seen as a kind of exoteric information, which laypeople may use to judge the relative trustworthiness of different groups of experts.

Let me now add my own two cents. I want to suggest that we view epistemic humility as yet another type of exoteric information in an expert's discourse. When seen through this lense, the Humility Heuristic becomes a heuristic about how to incorporate a particular kind of exoteric information. To take a simple example, let's suppose that you're confronted with a disagreement between two medical doctors about the effects of cannabis on clinical depression. Doctor A believes that cannabis *is* an effective treatment of depression, whereas Doctor B believes it *isn't*. Furthermore, let's suppose that you know (perhaps from a previous encounter) that Doctor A has expressed epistemic humility about a different medical issue (say, about the effects of musical treatment on epilepsy), whereas Doctor B *hasn't* (at least not to your knowledge). Given this, the Humility Heuristic says that you should treat this fact as a *pro tanto* reason to trust Doctor A more than Doctor B when it comes to judging the effects of cannabis on clinical depression.

Needless to say, there might be other, potentially more weighty, reasons to think that Doctor B is more trustworthy than Doctor A. Perhaps a third expert has appraised Doctor B, but not Doctor A. Or perhaps Doctor B's past track-record is more impressive than Doctor A's. The Humility Heuristic doesn't say anything about how to incorporate these other kinds of exoteric information. It just says that you should treat the fact that Doctor A has expressed epistemic humility about the effects of musical treatment on epilepsy as at least a weak *pro tanto* reason to think that Doctor A is more likely than Doctor B to be on the right side of the disagreement about the effects of cannabis on clinical depression.

So far, I've focused on cases where two or more experts disagree. Often, however, a layperson will receive testimony from only a single expert. In such cases, the layperson still faces a challenge of determining how much trust to place in the expert's testimony. After all, not all putative experts are genuine experts, and it can often be difficult to tell who is who. A particularly salient example of this arises from a phenomenon that Ballantyne (2019) and Gerken (2018) call "epistemic trespassing:" roughly, the phenomenon of experts testifying outside their area of expertise. Here is a real-world example, which Ballantyne uses to illustrate the general phenomenon:

> Linus Pauling, the brilliant chemist and energetic proponent of peace, won two Nobel Prizes—one for his work in chemistry, and another for his activism against atomic weapons. Later, Pauling asserted that mega-doses of vitamin C could effectively treat diseases such as cancer and cure ailments like the common cold. Pauling was roundly dismissed as a crackpot by the medical establishment after researchers ran studies and concluded that high-dose vitamin C therapies did not have the touted health effects. Pauling accused the establishment of fraud and careless science. This trespasser did not want to be moved aside by the real experts. (Ballantyne 2019, p. 367)

This kind of epistemic trespassing is all too familiar. What can we do to tell who is engaging in epistemic trespassing, and who is not?

Once again, I want to suggest that the Humility Heuristic can provide part of the answer here. If you know that a given expert has (perhaps on a previous occasion) declined to testify outside his or her area of expertise, the Humility Heuristic says that you should treat this fact as at least a weak *pro tanto* reason to trust the expert on this occasion. Of course, we don't always have access to information about whether a given expert has declined to testify outside his or her area of expertise. But when we *do*, the Humility

Heuristic allows us to use this information as a basis (albeit a defeasible one) on which to distinguish cases of genuine expert testimony from cases of epistemic trespassing.

6. Conclusion

My aim in this paper has been to propose a novel heuristic to help guide our search for trustworthy advisors. In slogan form, the heuristic says: *people worth trusting admit to what they don't know*. I argued that this "Humility Heuristic," suitably precisified, offers accurate guidance in a wide range of ordinary situations. The qualification "in a wide range of situations" has been left deliberately vague. The question of how often, exactly, the Humility Heuristic will provide accurate guidance ultimately depends on the kinds of epistemic situations that we find ourselves in. But even if I've been too optimistic in my assessment of the heuristic, I hold out hope that a better understanding of the conditions under which it *does* provide accurate guidance may prove useful in determining when to rely on it, and when not.

Appendix: Proof of Sufficiency Result

Sufficiency Result: The Humility Heuristic is accurate if the following conditions obtain:

C1 $P(Tp|{\sim}p \,\&\, Hq) < P(Tp|{\sim}p)$

C2 $P(p|Hq) \geq P(p)$

C3 $P(Tp|Hq) \geq P(Tp)$

Proof: By Bayes' Theorem, C1 is equivalent to:

$$P({\sim}p \,\&\, Hq|Tp)P(Tp)/P({\sim}p \,\&\, Hq) < P({\sim}p|Tp)P(Tp)/P({\sim}p) \quad (1)$$

By the Ratio Formula, C2 is equivalent to:

$$P({\sim}p \,\&\, Hq) \leq P({\sim}p)P(Hq) \quad (2)$$

From (1) and (2), it follows that:

$$P({\sim}p \,\&\, Hq|Tp)/[P({\sim}p)P(Hq)] < P({\sim}p|Tp)/P({\sim}p) \quad (3)$$

By the Ratio Formula, (3) is equivalent to:

$$P({\sim}p \,\&\, Hq \,\&\, Tp)/[P(Tp)P(Hq)] < P({\sim}p|Tp) \quad (4)$$

By the Ratio Formula, C3 is equivalent to:

$$P(Tp)P(Hq) \leq P(Tp \,\&\, Hq) \tag{5}$$

From (4) and (5), it follows that:

$$P(\sim p \,\&\, Hq \,\&\, Tp)P(Tp \,\&\, Hq) < P(\sim p|Tp) \tag{6}$$

By the Ratio Formula, (6) is equivalent to:

$$P(\sim p|Tp \,\&\, Hq) < P(\sim p|Tp) \tag{7}$$

Since $P(\sim p|\cdot) = 1 - P(p|\cdot)$, (7) is equivalent to the Humility Heuristic. ■

Acknowledgments: Earlier versions of this paper were presented at Aarhus University and University of Southern Denmark. I'd like to thank the audiences on those occasions for helpful comments and questions. Thanks also to Mark Satta and two anonymous reviewers at *Social Epistemology* for valuable feedback.

References

Ballantyne, N. (2019): "Epistemic Trespassing." In: *Mind* 128, pp. 367–95.

Battaly, H. (2008): "Virtue Epistemology." In: *Philosophy Compass* 3, pp. 639–63.

Bovens, L. & S. Hartmann (2003): *Bayesian Epistemology*. Oxford: Oxford University Press.

Cassam, Q. (2016): "Vice Epistemology." In: *The Monist* 99, pp. 159–80.

Dellsén, F. (2018): "When Expert Disagreement Supports the Consensus." In: *Australasian Journal of Philosophy* 96, pp. 142–56.

Dorst, K. (2019): "Evidence: A Guide for the Uncertain." In: *Philosophy and Phenomenological Research*. Online first.

Elga, A. (2016): "Bayesian Humility." In: *Philosophy of Science* 83, pp. 305–23.

Fallis, D. (2009): "What Is Lying?" In: *The Journal of Philosophy* 106, pp. 29–56.

Frankfurt, H. (2005 [1986]): *On Bullshit*. Princeton, NJ: Princeton University Press.

Gerken, M. (2018): "Expert Trespassing Testimony and the Ethics of Science Communication." In: *Journal of General Philosophy of Science* 49, pp. 299–318.

Goldberg, S. (2010): *Relying on Others: An Essay in Epistemology*. Oxford: Oxford University Press.

Goldman, A. (2001): "Experts: Which Ones Should You Trust?" In: *Philosophy and Phenomenological Research* 63, pp. 85–111.

Greaves, H. and D. Wallace (2006): "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility." In: *Mind* 115, pp. 607–632.

Hardwig, J. (1985): "Epistemic Dependence." In: *The Journal of Philosophy* 82, pp. 335–49.

Joyce, J. (1998): "A Nonpragmatic Vindication of Probabilism." In: *Philosophy of Science* 65, pp. 575–603.

Kitcher, P. (1990): "The Division of Cognitive Labor." In: *Journal of Philosophy* 87, pp. 5–22.

Kitcher, P. (1993): *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*. Oxford: Oxford University Press.

Lackey, J. (2008): *Learning From Words: Testimony as a Source of Knowledge*. Oxford: Oxford University Press.

MacFarlane, J. (2011): "What Is Assertion?" In *Assertion*, J. Brown & H. Cappelen (eds.), Oxford University Press.

O'Connor, C. and J. Weatherall (2019): *The Misinformation Age: How False Belief Spread*. Yale University Press.

Schoenfield, M. (2014): "Permission to Believe: Why Permissivism is True and What It Tells Us About Irrelevant Influences on Belief." In: *Noûs* 2014, pp. 193–218.

Schultheis, G. (2018): "Living on the Edge: Against Epistemic Permissivism." In: *Mind* 17, pp. 863–79.

Stokke, A. (2016): "Lying and Misleading in Discourse." In: *The Philosophical Review* 125, pp. 83–134.

Titelbaum, M. (forthcoming): *Fundamentals of Bayesian Epistemology*. Oxford: Oxford University Press.

Whitcomb, D., H. Battaly, J. Baehr, and D. Howard-Snyder (2015): "Intellectual Humility: Owning Our Limitations." In: *Philosophy and Phenomenological Research*, 94, pp. 509–39.

White, R. (2005): "Epistemic Permissiveness." In: *Philosophical Perspectives* 19, pp. 445–59.