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

Mattias Skipper
Draft of May, 2020

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"—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, provide a Bayesian argument for it, and demonstrate its practical worth by showing how it can help address a number of 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.

The challenge is a non-trivial one. 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. 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") to help guide our search for trustworthy advisors. In slogan form, the heuristic says:

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

I'll give this heuristic a precise probabilistic interpretation (§2), provide a Bayesian argument for it (§3), defend it against some possible worries (§4), and demonstrate its practical worth by showing how it can help address a number of difficult challenges in the relationship between experts and laypeople (§5).

2. The Humility Heuristic in Probabilistic Terms

The first task is to sharpen the proposal. We 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.

To analyze this situation in a precise manner, a bit of formal machinery will be helpful. Let P be the (unique) *rational credence function* of the advisee prior to

[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.

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. I'll make three standard assumptions about *P*.[2] First, I'll assume that *P* obeys the 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). 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. Throughout most of the paper, 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 way of *asserting the truth of p*. 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).[3] 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: rather than saying "I don't know," the advisor might just as well say "I couldn't tell you" or "I'll have to owe you an answer on that one."[4] Let's say that an agent who admits to not knowing whether a given proposition is true thereby

[2] 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.

[3] 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).

[4] Note, in particular, that nothing is going to turn on whether the advisor admits to lacking (i) *knowledge* or (ii) *justification to believe*. Rather than saying "I don't know," the advisor might just as well say "I don't have sufficient evidence to answer that question." However, as it is much more common in ordinary discourse to talk about what we do/don't *know* than to talk about what we do/don't *have justification to believe*, I'll use the locution "I don't know" as the paradigmatic way of expressing the kind of epistemic humility that I'm interested in.

expresses *epistemic humility* about that proposition; and let's write "*Hp*" to denote the proposition "the advisor expresses epistemic *H*umility about *p*."[5]

Needless to say, there are many other answers that one might give in response to a question of the form "Is *p* true?" For example, rather than outright asserting *p*, one might express a weaker kind of commitment to the truth of *p* by saying things like "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'd like to keep matters relatively simple by restricting attention to the two answers described above.

With these preliminaries in hand, we're ready for the official statement of the Humility Heuristic (where *p* and *q* are arbitrary propositions):[6]

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

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* should be lower than the advisee's credence in *p* given that the advisor testifies to *p and* admits to not knowing whether *q* is true. That's

[5] Two further remarks on terminology. First, the term "epistemic humility" (together with its close cousins like "epistemic modesty" and "intellectual humility") 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 whether your beliefs are rational. My use of the term "epistemic humility" is different from both Elga's and Dorst's. Note, however, that all three notions are introduced as (semi-)technical terms, not competing analyses of the same intuitive concept. For a discussion of what is involved in our ordinary thought and talk about intellectual humility, see Whitcomb et al. (2017).

Second, the term "trust" has likewise been given a number of different meanings in the literature. In particular, there is an ongoing debate about how best to capture our ordinary understanding of what it means to trust someone, and what it means to be worthy of being trusted; see, e.g., Baier (1986), Hawley (2014) and Nguyen (forthcoming). Again, however, my understanding for present purposes of what it means for a person to be trustworthy (although, I take it, not entirely divorced from our ordinary conception of trustworthiness) will be stipulative: you have reason to trust a person on a given occasion iff you have reason to think that the advisor speaks the truth on that occasion.

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

the precise meaning of the slogan "people worth trusting admit to what they don't know."

Before I present my argument in favor of the Humility Heuristic, let me clarify a few points about what the heuristic says, and why we should care to provide an argument for it in the first place.

The first thing to note is 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 more intuitive terms: 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; compared, that is, to people who *don't* admit to what they don't know. In this respect, the Humility Heuristic is a relatively weak claim, and it's natural to wonder whether it may be strengthened in various ways. I'll briefly return to this point in §4, but a detailed investigation must wait for another occasion. In this paper, the focus will be on establishing the purely ordinal claim.

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

$P(p|Tp) > P(p)$
The fact that the advisor testifies to p is evidence for p.

$P(p|Hq) > P(p)$
The fact that the advisor expresses humility about q is evidence for p.

Neither inequality follows from the Humility Heuristic. In fact, the Humility Heuristic may be accurate even if neither Tp nor Hq supports p.[7] 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.

[7] Here is a quick proof by counterexample: 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.

Third, 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.[8] 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. And, as I'll argue below, I think this question can be given a positive answer.

A final point before we proceed. In defending the Humility Heuristic, I don't take myself to take a stance on a controversial issue. I suspect that many readers will be sympathetic to the Humility Heuristic even before working through the details of the argument presented below. Nevertheless, I believe that there is something to be gained from working through those details. As I see it, it can often be interesting and illuminating to search for a theoretical justification or vindication of a thesis, even if that thesis is presumed to be true at the outset.[9] That's the spirit in which the following investigation is to be taken.

3. A Bayesian Argument for the Humility Heuristic

The backbone of the argument is the following formal result:

> **Sufficiency Result:** The Humility Heuristic is accurate provided that the following conditions obtain:
>
> C1: $P(Tp|{\sim}p \,\&\, Hq) < P(Tp|{\sim}p)$
>
> C2: $P(p|Hq) \geq P(p)$
>
> C3: $P(Tp|Hq) \geq P(Tp)$

This result is simply a theorem of the probability calculus (see the Appendix for a proof). But it holds valuable information about the conditions under which the

[8] 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.

[9] Compare: even if you already accept Probabilism and/or Conditionalization as norms of credence, it might still be worth your while to investigate whether those norms can be justified or vindicated on purely accuracy-based grounds (cf. Joyce 1998; Greaves and Wallace 2006).

Humility Heuristic is accurate: it tells us that the Humility Heuristic is accurate whenever a certain set of conditions, C1-C3, obtain. The question, then, is when these conditions obtain. Below I go over the conditions one by one, explaining what they say, what role they play in establishing the Sufficiency Result, and why we should expect them to obtain in most (but not all) contexts. As we'll see, there are some worries one might have about each of the conditions, as well as about the argument as a whole. I'll address some of these as we go along, but I'll defer the worries I take to run a bit deeper until §4, when the positive case for the Humility Heuristic is on the table.

Remarks on C1: The first condition is also the most crucial 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 *one* proposition is true will (other things being equal) also be more likely 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. In other words, the fact that a person expresses epistemic humility about q is typically going to be at least a weak *pro tanto* reason to think that the person wouldn't assert p, if p were false.[10]

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

[10] 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 this question in §4.1.

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 to these questions, the foreign minister provides the following two answers:

A1: "I'm afraid we don't know enough to answer that question."

A2: "Yes, it would."

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? This, as always, depends on your background evidence. But presumably, on most realistic ways of filling in the details of the case, you should 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.[11] That's all it takes for C1 to obtain.

I submit that most ordinary situations are like Press Conference in this respect. Typically, it's reasonable to treat the fact that a person expresses epistemic humility about some 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

[11] 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.

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 (or perhaps only a miniscule reason[12]) to think that your friend won't lie about who invented the light bulb.

However, even if C1 isn't immune to counterexamples, it may 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.

Remarks on C2: 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*.

The reason why C2 is needed for the Sufficiency Result is simply that, in cases where *Hq* is direct evidence against *p*, *Tp* & *Hq* can fail to be stronger evidence for *p* than *Tp* alone, 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*, had *p* been true).[13] 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*.

[12] I add this qualification because, on closer inspection, it's not entirely clear that you shouldn't become at least *slightly* more confident that your friend won't lie about who invented the light bulb. Still, with enough ingenuity, I suspect it's possible to construct a genuine counterexample to C1 along these lines.
[13] This example is inspired by Goldberg's (2010, ch. 6) discussion of inferences from "absence of evidence" to "evidence of absence."

After all, if *p* had been true, your friend would most likely have known *q*, in which case they most likely wouldn't have admitted to not knowing whether *q* is true. Thus, assuming that *Hq* is a strong enough rebutter of *p,* we have a case where *Tp* & *Hq* doesn't support *p* more strongly than *Tp* alone, contradicting the Humility Heuristic.

But again, what matters for present purposes is whether C2 *typically* obtains; which 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 seems to cover a quite wide range of ordinary cases: the fact that your colleague admits to not knowing whether the Lakers beat the Celtics last night seems to have no (or at least only a minuscule) 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 (or at least only a minuscule) evidential bearing on whether the chemical structure of water is H_2O; and so on. More generally: unless you have a special reason to think that the question of whether your advisor knows *q* has a direct evidential bearing on whether *p* is true, C2 will (*a fortiori*) obtain.

Remarks on C3: 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*.

The reason why C3 is needed for the Sufficiency Result is a little more subtle: in cases where *Hq* is evidence against *Tp*, the Humility Heuristic can fail to be accurate, even if C1 and C2 both obtain. Here's an example: suppose 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 knowledge of what people tend to

know 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, contrary to C3, 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 strong evidence (for you) that your friend doesn't know the capital of France either.

We can then 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. Here is a case, then, where C3 fails to obtain, and where, as a consequence, the Humility Heuristic fails to be accurate.

But, once again, what matters is whether C3 *typically* obtains; and, once again, I think it does. The reasons are similar to those offered in favor of C2. At the very least, C3 will obtain whenever *Hq* is evidentially irrelevant to *Tp* (relative to the advisee's background evidence), and this seems to cover a quite wide range of ordinary cases: the fact that your mother admits to not knowing who founded Marlboro seems to have no (or at least only a minuscule) 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 (or at least only a minuscule) evidential bearing on whether she will tell you that today's meeting is cancelled; and so on. More generally: unless you have a special reason to think that the

question of whether your advisor knows *q* has a direct evidential bearing on whether your advisor will testify to *p*, C3 will (*a fortiori*) obtain.[14]

So far, so good. We've now seen that the Humility Heuristic is accurate whenever C1-C3 obtain; and we've seen that these conditions obtain in a wide range of ordinary situations. But what about when they *don't* obtain? Is the Humility Heuristic inaccurate in all such cases? No. Just as none of the conditions is individually *sufficient* for the Humility Heuristic to be accurate, none of them is individually *necessary* either. In fact, it turns out that the strongest logical combination of C1-C3, which is necessary for the Humility Heuristic to be accurate, is their *disjunction*. That's our next result:

> **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.[15] It tells us that the Humility Heuristic is guaranteed to be inaccurate when C1-C3 all fail to obtain at the same time. If the foregoing remarks are on the right track, we should of course expect such situations to be quite rare. However, there is a different result in the vicinity, which promises wider applicability:

---

[14] There is a slight complication here: 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, if so, doesn't this generate a broad class of counterexamples to C3? Perhaps so. But the relevant class of counterexamples to C3 wouldn't carry over as counterexamples to the Humility Heuristic. When a counterexample to C3 constitutes a counterexample to the Humility Heuristic, it's 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*.

[15] The proof of this result (and the next) is similar to the proof of the Sufficiency Result, and the details are left out.

Equivalence Result: The Humility Heuristic is equivalent to C1 provided that the following conditions obtain:

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*. Since C2* and C3* are logically stronger than C2 and C3, respectively, they will (in a trivial sense) obtain less often. Nevertheless, I think C2* and C3* will obtain in a fairly wide range of ordinary situations. As suggested in the foregoing remarks on C2 and C3, it often seems reasonable to assume that *Hq* has no (or at least only a minuscule) 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 viewed as the most crucial condition.

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 two of the most interesting worries that have come to my attention. Ultimately, I don't think either worry has much force against the central thesis of the paper. But they each raise important questions about the scope and limitations of the Humility Heuristic 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 needn't be trustworthy on matters of developmental psychology; someone who is trustworthy on matters of English literature needn't be trustworthy on matters of US foreign politics; and so on. More generally, someone who is trustworthy in one domain needn't be trustworthy in other, far removed domains. Doesn't this suggest that we should only expect the Humility Heuristic to be accurate when *p* and *q* fall within the same domain, or at least suitably similar domains?

There is clearly something right about the observation that people's degree of trustworthiness varies from domain to domain. One can, of course, 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. And 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 for present purposes, then, is whether this purely ordinal claim is true (or rather *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 ~*p* & *Hq* than given ~*p* alone. Is this condition satisfied even if *p* and *q* fall within very different domains? In particular: is it satisfied even if the advisor is much less trustworthy relative to the "*p*-domain" than relative to the "*q*-domain"? Given that the remarks on C1 in §3 are correct, the answer is positive: 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. To be sure, this is not to say that the fact that the advisor admits to not knowing whether *q* is

true raises their degree of trustworthiness relative to the *p*-domain by a *large amount*. The claim is just that the fact that the advisor admits to not knowing whether *q* is true makes it at least *slightly* less likely that the advisor will make false assertions about matters within the *p*-domain, including about *p* itself.

Consider an example: suppose (as seems reasonable) that you 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, yes: once again, 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 their 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.

In sum: 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. 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 more substantial about when expressions of epistemic humility have the most epistemic value, that is, when the difference between $P(p|Tp \,\&\, Hq)$ and $P(p|Tp)$ is most significant. If *that* was our goal, we'd do well to pay attention to the specific content of *p* and *q*. In particular, we'd do

well to pay attention to whether *p* and *q* fall within suitably similar domains. But that's a bridge we'll have to cross when we get there.

4.2. Hedged Assertions

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

> The Humility Heuristic, as stated, doesn't say anything about whether *p* and *q* must be *distinct* propositions. Yet, the Humility 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 presumably 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. For example, if I say "I believe she's gonna make it, but I might be wrong" this will (at least typically, if not always) give you *less* of a reason to believe that she's gonna make it than if I say outright "I believe she's gonna make it." Doesn't this impose quite significant limitations on the scope of the Humility Heuristic?

I think this worry is basically sound… exception for one key point: the logical form of a hedged assertion is not well captured by the conjunction "*Tp* & *Hp*." As we recall from §2, 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.[16] As such, the Humility Heuristic was never supposed to say anything about hedged assertions in the first place. In particular, it doesn't say that the hedged assertion "I believe she's gonna make it, but I might be wrong" gives you more reason to believe that she's gonna make it than the outright assertion "I believe she's gonna make it."

Nevertheless, I find the worry under consideration illuminating, because it reminds us to be careful about how we go about *generalizing* the Humility Heuristic, should we want to do so down the road. In particular, it teaches us that we can't straightforwardly generalize the Humility Heuristic to cover hedged assertions without running into a broad class of counterexamples.

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

Although the central aim of this paper is to lay the theoretical foundations of the Humility Heuristic, I'd like to close by looking in more detail at a specific application of the heuristic, in the hope of demonstrating the practical significance of what has been said so far. I've chosen to focus on a set of issues related to *expert testimony*. There are, no doubt, many other potential applications of the Humility Heuristic that would deserve a separate discussion, but I hope that the following considerations touch on challenges that many readers will be able to recognize from their own epistemic lives.

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 among themselves. When they don't, it can be difficult for the rest of us to figure out who's on the right side of the debate. After all, we, as laypeople, usually aren't in a position to adjudicate expert disagreements by looking at the relevant first-order evidence and

[16] For further discussion of the role of hedged assertions in ordinary discourse, I refer to Benton and van Elswyk (2020).
[17] See Hardwig (1985) and Kitcher (1990; 1993) for some excellent discussions on this point.

arguments ourselves. We simply don't have the requisite knowledge and competencies to do so. What can we do to help?

In his seminal discussion of this question, Goldman (2001, p. 94) introduces what I think is 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 is more readily accessible to the layperson. Needless to say, this distinction between esoteric and exoteric information isn't sharp; it admits of degrees, just like the distinction between "expert" and "layperson." But for present purposes, it won't hurt to talk about esoteric and exoteric information in categorical terms.

The central lesson of Goldman's discussion, then, is that, even if laypeople can't rely on esoteric information to adjudicate expert disagreements—that is, even if they aren't in a position to judge the bearing of the first-order evidence and arguments put forward by the experts—they might still be able to rely on various kinds of exoteric information to make an informed judgment about which expert is most worthy of being trusted. This raises a further question: what kinds of exoteric information do laypeople typically have at their disposal? Goldman himself discusses five broad categories of exoteric information related to, among other things, "dialectical superiority," past "track-records," and appraisals by "meta-experts." I won't go into detail with these here. Let me instead mention a different kind of exoteric information, which has been brought to my attention by Dellsén (2016).

Dellsén argues that the fact that there is disagreement among a group of experts on a given issue constitutes a *pro tanto* reason for the laypeople among us to trust the group of experts on issues on which they *agree*. For example, if I learn that a group of cosmologists disagree among themselves about whether the universe has a flat or curved geometry, I may treat this fact as a *pro tanto* reason to trust their consensus, if there is one, on the age of the universe. As Dellsén puts it: "expert disagreement

supports the consensus." One can, of course, take issue with this claim. But if it's right, it shows something interesting 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.[18]

Let me now add my own two cents: the suggestion I want to make here is 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, suppose 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 that it *isn't*. Suppose also that you know (perhaps from a previous encounter) that Doctor A has expressed epistemic humility about a different medical issue—about, say, the effects of musical treatment on epilepsy—whereas Doctor B *hasn't* (to your knowledge) expressed epistemic humility about this other medical issue. Given this, the Humility Heuristic tells you to treat this fact as a reason to think that Doctor A is more likely than Doctor B to be right about the effects of cannabis on clinical depression.

Of course, 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.

[18] I should note that Dellsén doesn't frame his proposal as one concerning exoteric information, but I hope that he will nevertheless be sympathetic to the spirit in which his proposal is put to use here.

What about cases where a layperson receives testimony from a *single* expert who isn't (to the layperson's knowledge) in disagreement with any other expert on the relevant issue? In such cases, we still seem face a challenge of determining how much trust to place in the expert's testimony. After all, not all *alleged* experts are *genuine* experts, and it can be difficult to figure out who is who. A particularly salient example of this comes from a phenomenon that Ballantyne (2019) and Gerken (2018) call *epistemic trespassing*: roughly, the phenomenon of experts testifying outside their area of expertise. Consider the following real-world example, which Ballantyne uses to illustrate the 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, I take it, a fairly widespread phenomenon. How can we tell when someone is engaging in epistemic trespassing, and when not?

Once again, I want to suggest that the Humility Heuristic can provide part of the answer: 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 tells you to 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 on previous occasions. But when we *do* have access to such information, the Humility Heuristic says to use it as a basis (albeit a fallible and defeasible one) on which to distinguish cases of genuine expert testimony from cases of epistemic trespassing.

6. Conclusion

We live in a world of less than fully reliable advisors. My aim in this paper has been to offer some guidance about how to navigate this predicament. More specifically, I've proposed a heuristic—I called it the "Humility Heuristic"—which says (in slogan form) that *people worth trusting admit to what they don't know*. I've argued that this heuristic, suitably precisified, provides accurate guidance in a wide range of ordinary situations. I've left the qualification "in a wide range of situations" deliberately vague. Ultimately, the question of how often, exactly, the Humility Heuristic will provide accurate guidance is going to depend on the kinds of epistemic situations that we find ourselves in. But even if it turns out that I've been too optimistic in my assessment of how often the Humility Heuristic will provide accurate guidance, I hold out hope that a better understanding of the conditions under which the Humility Heuristic *does* provide accurate guidance may prove useful in determining when to rely on the heuristic, and when not.

Appendix: Proof of Sufficiency Result

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:

$$P(Tp)P(Hq) \leq P(Tp \,\&\, Hq) \quad (5)$$

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

$$P(\sim p \,\&\, Hq \,\&\, Tp)P(Tp \,\&\, Hq) < P(\sim p|Tp) \quad (6)$$

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

$$P(\sim p|Tp \,\&\, Hq) < P(\sim p|Tp) \quad (7)$$

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

References

Baier, A. (1986): "Trust and Antitrust." In: *Ethics* 96, pp. 231–60.

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

Benton, M. and P. van Elswyk (2020): "Hedged Assertion." In: *The Oxford Handbook of Assertion*, S. Goldberg (ed.). Oxford: Oxford University Press. pp. 245–63.

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

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.

Hawley, K. (2014): "Trust, Distrust and Commitment." In: *Noûs* 48, pp. 1–20.

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.

Nguyen, T. C. (forthcoming): "Trust as an Unquestioning Attitude." In: *Oxford Studies of Epistemology*.

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

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 & D. Howard-Snyder (2017): "Intellectual Humility: Owning Our Limitations." In: *Philosophy and Phenomenological Research* 94, pp. 509–39.