



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/241239/>

Version: Accepted Version

Article:

Williams, R. (Accepted: 2026) epistemic charity examined. *The Philosophical Quarterly*.
ISSN: 0031-8094 (In Press)

This is an author produced version of an article accepted for publication in *The Philosophical Quarterly*, made available via the University of Leeds Research Outputs Policy under the terms of the Creative Commons Attribution License (CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Epistemic Charity Examined

Abstract

Epistemic charity says that we should (or may) interpret others in the most favourable way consistent with the evidence. Though this could be refined in several ways, this paper develops an argument against all of them, insofar as they endorse a certain “signature” of charity, where the favourable interpretation is combined with an acknowledgement of, and uncertainty over, countervailing evidence. Though initially plausible, the signature is susceptible to a version of triviality arguments prominent in the literature on conditionals. The second half of the paper evaluates responses to the argument and generalizations of it to other default epistemic rules. Ultimately, a resolution is suggested, requiring that the friend of epistemic charity combine their optimism with realism in a different way, via fragmented states of mind.

Wordcount 8,700

Keywords: Charity, Interpretation, Conditionals, Triviality, Fragmentation, Mindreading.

A. Epistemic charity, and the argument, in brief.

I want people to think that I am well-motivated, and not making stupid mistakes. However, I know that sometimes I act badly, and sometimes make dumb errors. The need for approval shows up in this way: in unclear cases I want people give me the benefit of the doubt. And when friends are involved, I do more than want charitable interpretation. I expect it. If friends don’t assume—as far as the evidence allows—that I’m acting out of good motives and not making dumb mistakes, that strikes me as a modest betrayal. What I expect for myself I feel obliged to extend to others. If, when

evidence is unclear, I did not give my friends the benefit of the doubt as regards motive and belief, I'd feel disloyal.

This paper develops an argument that a certain “signature” of charity is incompatible with very minimal constraints on rational belief and belief change. The argument will apply to many different flavours of epistemic charity (global or local, permissivist or mandatory, structural or substantive, defeasible or indefeasible) just so long as they incorporate the signature. Indeed, though charity is my focus in this paper, a form of this argument threatens epistemic structures of “default reasonable” defeasible assumptions, whatever their content.

In the subsections immediately below, I will canvass varieties of epistemic charity, present the signature of epistemic charity, and outline an argument against any principle that endorses that signature. In the second section, I regiment the argument and its premises and presuppositions more carefully. In the third section, I will evaluate the potential significance of the argument—from a way of rescuing charity at considerable metaphysical cost, to implications for the metaphysics of representation, to wider ramifications of the schematic argument in other settings. With the stakes set out, the fourth and final section will sketch what I take to be the best response to the puzzle.

A1. The varieties of charity.

My desire to be interpreted favourably could be a description of my personal idiosyncrasies or biases—a matter of descriptive psychology, and irrelevant to normative epistemology. But for today, the hypothesis is that my feelings are apt: one *can* and perhaps *ought* to interpret others charitably. This thesis can take many forms, all of which are threatened by the argument here.

A first dimension of variation concerns those to whom charity is owed. For example, Sarah Stroud (2006) has argued that part of the “special obligations” of friendship include what she calls the “proper epistemic bias of a good friend”. The bias involves, *inter alia*, seeking out some interpretation of one’s friend that squares the evidence at hand with a favourable evaluation of the friend’s character and actions. Being disposed to think the best of friends is, for Stroud, constitutive of *being a friend* (if you utterly lack the disposition, you are not really friends at all). But it is also a *demand* of friendship, something friends owe to one another. One might extend the Stroudian idea to other significant relationships; a communitarian might make a case for charity towards members of one’s family or tribe. Charity also seems at home in various institutional relationships: teacher to student or juror to accused (in some cases, we might identify charity in the rules that govern institutional roles: “Innocent until proven guilty”). What is key to the Stroudian class of principles of charity is the partiality of the normative thesis: charity is triggered by appropriate significant relationships between interpreter and interpretee. A contrasting approach is the long tradition, associated in particular with Davidson, that holds that a demand for charitable interpretation is universal.¹ Stroud argues that demands of charity are part of the “phenomenology” of friendship, and that we are inclined to view friends more charitably than strangers. The universalist can acknowledge all this, while insisting that we *should* interpret a stranger just as charitably as we do friends in the same evidential situation, and that friendship (say) just allows us to feel more easily the demands that are there all the time.

A second dimension of variation concerns the normative force of the charity principle. A principle of charity might make charitable interpretation obligatory (/when triggered), or merely permissible (/when triggered). A permissibility-focused principle of charity allows it to be a matter of personal character whether one adopts a charitable or cynical stance in interpreting others. Maybe it is an

¹ See the essays collected in Davidson (2001).

admirable character trait to be charitable across the board, or maybe it is not—compare what we might think about varying risk-attitudes in decision making. This variation in force may interact with variation in the recipient of charity. For example, one might think that a universal charity permissivism is compatible with a Stroudian thesis about obligation to be charitable to friends. If we go that way, we would think that charity, while always permissible, becomes in suitable circumstances, obligatory.

A third dimension of variation concerns the kind of favourability that charity seeks to maximize. Charity could tell us to pick an interpretation that makes the target most *intelligible*, or most *structurally rational*, or most *justified and moral*.² Again, interactions are possible: there could be universal obligatory demands to maximize structural rationality, universal permission to maximize justification and morality, with this turning into obligation only when triggered by appropriate underlying relationships such as friendship.

The final dimension of variation concerns the defeasibility of any obligation to be charitable. Consider the following puzzle situation: I know having seen blurred CCTV footage that someone viciously assaulted Victim, and I know that either A or B did the deed. A and B are fit recipients of charity (perhaps they are friends of mine). Since I cannot tell who was pictured, I do not take myself to have compelling evidence that A did the deed. So this is a scenario where charity applies: I can give A the benefit of the doubt, or not. The trouble is if I should, in light of charity, believe A did

² For one early discussion of variations of principles of charity see Grandy (1973). For another see the role of inductive methods in Lewis (1974). Some more recent variations on charity include a central role for ascribing knowledge (Williamson (2000)) and a central role for maximizing morality: Williams (2018). See also the discussion in chapter 1 of structural vs. substantive rationality maximization in my (Williams 2020). Lewis, Williamson and myself are all working with metaphysical rather than epistemic principles of charity, but as will be discussed later, the space of options here is similar.

not do it, by parity of reasoning it looks like I should, in light of charity, believe that B did not do it. But now I have ended up with three beliefs of the form: not p, not q, but: p or q. A friend of charity may wish their account avoid predicting a commitment to this inconsistency. They might do so by insisting that an obligation to be charitable will be defeated in contexts where it is impossible coherently to be charitable to everyone. Differences in how to spell out this defeasibility and, more generally, how charity applies in multi-agent contexts is the final dimension of variation in epistemic principles of charity.

This paper develops an argument that a certain “signature” of charity is incompatible with very minimal constraints on rational belief and belief change. The argument will apply whether the principle of charity is global or local, permissivist or mandatory, structural or substantive, defeasible or indefeasible — so long as it predicts the signature.

The plan is as follows. In the subsections immediately below, I will present the signature of charity, and sketch an argument against any principle that endorses that signature thesis. In the second section, I regiment the argument and its premises and presuppositions more carefully. In the third section, I will evaluate the potential significance of the argument—from a way of rescuing charity at considerable metaphysical cost, to implications for the metaphysics of representation, to wider ramifications of the argument in other settings. With the stakes set out, the fourth and final section will sketch what I take to be the best response to the puzzle.

A2. The signature of charity.

Epistemic charity principles embody a kind of epistemic optimism. They break what would otherwise be an epistemic symmetry in favour of the more favourable interpretation of another. I say that this motivates what I call the “signature” of epistemic charity. I sketch an illustration,

formulate the abstract principle, and then tell you why friends of epistemic charity seem committed to the signature.

Consider the following case:

Sally was in charge of the control tower on a stormy night when a plane crashed into the mountains. The black box show that it followed a risky path near the mountains under instructions from the control tower. What's at issue in the investigation is whether Sally issued the instructions maliciously, seeking to take revenge on a former boss (who was on board the plane and perished in the crash), or whether she issued the instructions in the sincere belief that the route was less risky than flying higher through the storm. Already it has emerged beyond reasonable doubt that (i) Sally issued the instructions; that (ii) maliciously crashing the plane would be an evil act (whatever the boss did, motives of revenge could not justify killing all the innocents on board), and that (iii) CCTV recordings in the control tower could provide decisive evidence in favour of malice. Unfortunately, storm surges corrupted the files where the footage was stored. Restoring the footage is in progress, and the best version available so far suggests that Sally failed to monitor the storm path and plotting alternative routes—but it's glitchy and fuzzy enough that it's really too early to tell.

Let E be the proposition that Sally did not monitor the path of the plane at all in the control room. This is a proposition concerning possible behaviour in the control room on the night that, if known to be true, would decisively support the hypothesis that Sally acted out of malice. So rationality would require you to believe that Sally acted out of malice if you believed that E. Given the glitchy and fuzzy footage, however, you are uncertain whether E is the case. (Fill in the details of how glitchy and fuzzy the footage is as you wish, until you are happy to agree that uncertainty on this

matter is the appropriate reaction). There is therefore still reasonable doubt in this case. The signature of a principle of charity is that in an uncertain case like this, the interpreter is at least permitted to believe that Sally is non-malicious.

I call the pattern of attitudes we see in this case a signature of charity because all the various principles of charity can agree with it, or a close variant. On the versions that require we stand in a particular relation to Sally (friendship, tribal affiliation, or juror-accused) we simply add that detail to the vignette. Since what is obligatory is permissible, both permissivist and obligatory flavours of epistemic charity will agree on the claim that belief in Sally's non-maliciousness is permissible. Modest adjustment of the details can tailor it to versions of the principle of charity where the "favouring relation" among psychological states which charity promotes doesn't favour innocent over malicious psychologies but rather e.g. structurally rational over structurally irrational ones.

In abstract form, the signature of epistemic charity is the following. Take a situation where it is known that there are only two competing interpretations of a proper target of charity *x*: GOOD and BAD, and that *E* is a proposition describing what would be compelling evidence for BAD. It is then rationally permissible for a person to be uncertain whether *E* obtains, and nevertheless believe that GOOD is the correct interpretation of *x*, and (so) disbelieve that BAD is the correct interpretation of *x*.

All the various principles of epistemic charity *can* agree with the signature. *Should* they agree to it? I say: any friend of epistemic charity must say *something* to explain how their epistemic optimism gets tempered by realism. Sensible charitable interpreters believe the best of others (optimism), while holding themselves open to bad news about them (realism). The signature of charity is the *obvious* and *natural* way of doing this. On this conception of charity, one limits one's optimism to a circumscribed set of topics, e.g. Sally's psychology, and suspends judgement on other matters, such

as what physical movements are depicted on glitchy footage—standing ready to form a firmer opinion on this matter once the technicians clean up the tape. This circumscription of the optimism leads directly to the signature. Turning this around: imagine a version of epistemic charity that denied the signature. This would be uncircumscribed epistemic optimism, permitting us not only to think the best of Sally, but to prejudge every single bit of potentially decisive evidence against her (allowing uncertainty on even one proposition that would support the malicious hypothesis will produce an instance of the signature). Uncircumscribed epistemic charity seems like it permits us, or even obliges us, to wipe from our minds entirely the evidential symmetry of the situations in which charity is called for. That’s no part of the phenomenology of charitable interpretation as I experience it.

I haven’t argued that the signature is an inevitable part of epistemic charity. I am a friend of epistemic charity myself, and ultimately what I take away from the puzzle I’m about to describe is that we need a *different* conception of how the charitable interpreter tempers their optimism with realism—one which I describe in the final section. But in order to see why the (controversial, unobvious) conception of the psychology of charitable interpretation I describe at the end is necessary, we need to see why the natural, obvious way of doing the job won’t work—why the friend of charity can’t just accept the signature.

A3. The consequences of epistemic charity

Before describing the main argument against the signature of charity, I will give a warm up argument against it. This argument will have more loopholes than my main puzzle, but it’ll still show us interesting things about what epistemic charity must be like.

The signature of charity is a situation where we suspend on some evidence-describing proposition E, and disbelieve in some psychological attribution BAD, where E is compelling evidence for BAD. Let us suppose that we are charitable interpreters with that combination of attitudes. Now consider the conditional:

if E then BAD.

Since E describes compelling evidence for BAD, this conditional should sound great! Let us suppose, then, that we believe it.

This leaves us in the following situation: we believe a conditional, but as part of the signature of charity, we believe the negation of its consequent. From the conditional and the negation of the consequent, the negation of the antecedent, not-E, follows. So, apparently, we should believe not-E. However we were, on the contrary, suspending on whether E or not-E.

To turn this into an argument against charity, one needs first, the claim that we are epistemically *obliged* in relevant situations to believe the conditional if E, then BAD; and second, the claim that it's not epistemically permissible for someone to believe each of the premises of an (obviously, known-to-be) valid argument, and adopt an attitude *other* than belief to the conclusion. If you assume both of these, then we have a reductio of what I called the signature of charity. We may conclude: it is not permissible to suspend on E while believing that BAD is false.

This warm-up argument has two obvious vulnerabilities. First, the friend of epistemic charity may reject the conditional. By stipulation, E was good evidence (by the charity-lovers' lights) for BAD. That tells us only this: that if she were to learn that E as total evidence, she should believe BAD. To say that in such a circumstance she should believe a conditional is (while very plausible) a move

that involves a link between learning-dispositions and believing conditionals that, since David Lewis's work on triviality results in the 1970s, we know to be in the ballpark of paradox.

Second, the logical norm that this argument uses is one that an influential minority of epistemologists already reject for independent reasons. To present the parallel case:³

Painted mules:

Visiting the zoo on a day off, I see a stripey equine in a pen. I rationally believe: that's a zebra. I know a modicum of zoology, so I also believe: if that's a zebra, it's not a painted mule. But given my evidence is just a casual glance at a stripey equine, I would not be justified in believing: that is not a painted mule. So I permissibly suspend on that last question.

In this case, as in the above, we rationally believe each of the premises of an (obviously, known-to-be) valid argument, and suspend belief in its conclusion. So the logical norm above is violated. The friend of epistemic charity has a cluster of partners in crime here, and the argument against her will work only if we can defend this norm.

This warm-up argument shows us some interesting things about the epistemology of epistemic charity: that it either says some surprising things about when to believe conditionals, or says some surprising things about logical norms. The friend of epistemic charity may embrace these as features of their position. The main argument to follow does not allow such loopholes.

³ Cf. (Dretske 1970).

B. Regimenting the argument against epistemic charity.

The starting point of my main argument is an upper limit on how much benefit of the doubt one can properly offer. Where E is the proposition describing compelling evidence for the unfavourable interpretation BAD, one should conform to the following imperative:

If E, believe that BAD is correct!

Uncertainty over the antecedent may lead one to be unsure whether the imperative is triggered. But if one is trying to conform, one should at least satisfy the following derivative rule:

If one believes that E, then believe that BAD is correct!

Turning this into a claim about rational belief states: in any rational belief state, if one believes that E, then one believes that BAD is correct. This premise I call LIMIT.

I will present the main argument in a very general setting.⁴ I begin with assumptions about the ontology of possible doxastic attitudes one could take to a proposition. What the argument will assume is that there are a set of such attitude-types S, ordered by comparative confidence $>$. The attitude a subject takes to proposition p will be written $B(p)=s$, where s is one of the attitude types in S. And the *conditional* attitude a subject takes to a proposition p on the assumption that q will be written $B(p|q)=s$, where again, s is one of the attitude types in S. We also assume that there is a topmost, “most confident” belief-state-type. We label that most confident attitude type “1”, so that

⁴ Santorio and Williams (2022) develop this highly general regimentation of “triviality style” arguments in the tradition of (Lewis 1976).

$B(p)=1$ says that in the relevant sense (i.e. with the respect to the attitude-types in S), the subject is as confident a state with respect to p as they can be in towards any proposition.

These assumptions are very liberal. For all they say, there may be just three doxastic attitudes: belief, uncertainty, and disbelief (/conditional belief, conditional uncertainty, conditional disbelief), listed in descending order of comparative confidence. $B(p)=1$ will then be read as saying that the subject believes p . “Informal epistemology” often talks in this way, and the informal gloss on the argument above can be understood as presupposing that coarse-grained doxastic ontology. On the other hand, the assumptions just set out are also compatible with a standard Bayesian framework in which there are infinitely many doxastic attitudes, aka degrees of belief, representable by real numbers between zero and one. In such a setting, comparative confidence corresponds to the usual ordering on the reals. Further, the assumptions are compatible with more exotic representations of belief where there is no linear ordering on doxastic attitudes. An example is the imprecise credence model, where attitudes are represented by sets of real numbers between zero and one, where the singleton interval $[1]$ represents the topmost confidence, but where, for example, the attitudes represented by $[0.7,0.3]$ and $[0.8,0.2]$ are incomparable. In short, it will be difficult to escape the argument to follow simply by rejecting its implicit ontology of belief.

Possible belief states will be represented by overall assignments of attitudes in S to propositions (for the categorical attitudes) and pairs of propositions (for conditional attitudes). Many possible belief states are irrational, but the set R will be the set of possible belief states that satisfy every rational requirement. I’ll be neutral about what these rational requirements are—everyday epistemology will have one story, Bayesians another, more exotic formal epistemologies a third, and the total packages are a matter of controversy. Following Santorio and Williams (2022), we make three, very minimal, assumptions about R .

- The first assumption is IDENTITY. It is that in a rational belief state, one has the topmost level of confidence in a proposition, conditional on that proposition itself. That is, for all B in R, $B(p|p)=1$.
- The second assumption is BOUND, and relates conditional beliefs to categorical beliefs. It is that for all B in R, when $B(p|q)=1$, then it is not the case that $B(q)>B(p)$.
- The third assumption is CLOSURE. This says that for all B in R, there is a rational belief state B* whose *categorical* degrees of belief in each proposition p are exactly B's the *conditional* degrees of belief in p, given q

IDENTITY is pretty obvious, and I won't linger to defend it here.

BOUND is extremely plausible. Here are two ways of seeing that. First, when you are maximally confident in p given q, then by your lights, any q situation will be one where p is the case. So any confidence you have in q situations obtaining automatically transfers to same confidence in p-situations obtaining. It should not, therefore, be possible for your confidence to be lower in q than in p, in those circumstances. The second way of seeing its plausibility is more indirect, and involves two premises. (a) when you are maximally confident in p given q, you should have exactly the same level of confidence in the proposition that p and q as you have in the proposition that q (conjoining p "adds nothing" to q by your lights); (b) you shouldn't be *more* confident in a conjunction, p and q, than you are in one of its conjuncts, p. Put (a) and (b) together, and you have the thesis that you shouldn't be more confident in q than in p, which is all that BOUND says.

CLOSURE is motivated in the following way. Your conditional attitudes express the degree of belief you have in p on the assumption that q is the case. Now suppose you learn, with maximal

confidence, that q is the case, and q is all you learn. Then your rational, posterior, categorical beliefs should match your rational, prior, conditional-on- q beliefs. And the existence of a rational belief state satisfying this description is all that CLOSURE asserts. You can resist this motivation for CLOSURE, only if you could argue that it is literally *impossible* to learn q (as total information) with maximal confidence. For the instance of CLOSURE used in the argument below that is a very tall order.⁵

It's worth noting that we will only need single instances of CLOSURE, IDENTITY and BOUND to run the argument below, so it's no good producing *recherché* counterexamples to those principles in an attempt to block the argument, unless those undercut the particular instances used. For example, we'll only need to appeal to one instance of CLOSURE involving the particular evidence proposition E .

The three principles are compelling and it should be headline news if we have to give any one of them up. But if you are tempted to deny them, relative to your preferred ontology of belief, you may have to further rebut more local arguments for them within your preferred epistemology. For example, IDENTITY and BOUND follow directly from the standard ratio formula for conditional probability in the Bayesian setting, and CLOSURE follows from the assumption that rational update (on a proposition learned as total information) proceeds by conditionalization, at least for learnable q .

⁵ One lesson to take from the generalizations of charity here and in Santorio and Williams (op cit) is that different instances of the abstract puzzle call for different responses. When considering instances of the puzzle involving epistemic modals and conditionals, I'm attracted to the response whereby we deny closure. That's because where our own state of mind is caught up in the subject matter of a proposition, the very act of update changes the facts we are tracking. A rational response involves not just learning those facts (modelled by conditionalization), but also reflexive compensation for the changes induced by that learning. But indeterminacy or the psychology of other people are *not* cases where the subject matter is our own mind. There's no particular reason to think the same story will work (unless we take the radical step of endorsing expressivism about indeterminacy or others' psychology—something considered below).

The argument will also use a regimentation, using the present lingo, of the LIMIT assumption discussed earlier:

- LIMIT: for all belief states B in R , if $B(E)=1$ then $B(BAD)=1$.

The argument then runs as follows, for an arbitrary B in R :

1. $B(E|E)=1$ (IDENTITY)
2. $B^*(E)=1$ (where B^* is the R -belief state given by CLOSURE on E).
3. $B^*(BAD)=1$ (by LIMIT applied to rational belief state B^*).
4. $B(BAD | E)=1$ (by construction of B^*).
5. $\neg B(BAD)<B(E)$ (by BOUND)

This rules out, for all B in R , a drop in degree of belief from the evidence proposition E to the proposition that the bad interpretation is correct. And that's our result! For a permissible credence drop in this case, recall, is exactly the signature of an epistemic principle of charity.

The regimented argument, I have said, works with the set of all rational belief states R . However, the argument can be run schematically on parameter R . Tweaking it may be useful in responding to some concerns. Suppose, for example, that someone insists that the limit constraint linking E to BAD doesn't hold for all rational belief states whatsoever. They may insist that the link holds only relative to belief states of agents who start from a reasonable set of priors, or only for belief states of agents who don't possess defeaters for the inference from E to BAD . Let R^* then be the set of rational belief states that start from sensible priors and which lack defeating information.

IDENTITY and BOUND hold for all belief states in R^* if they hold for all in the wider set R .

CLOSURE on R^* is a stronger claim than closure on R , but unless the evidence-proposition E is

itself something that defeats the inference from E to BAD (which it isn't) and is rationally learnable for someone with sensible priors then CLOSURE on R^* will also hold. The conclusion of the argument, so construed, is that the signature of charity is not a rationally permissible belief state for someone with reasonable priors and no relevant defeaters. And that's just as much a reductio of the signature as the unrestricted conclusion would be.

C. Ramifications and extensions of the failure of epistemic charity

The immediate impact of the main argument, if sound, is to refute any epistemic principle of charity that entails the "signature" permissible confidence-drop. In this section I point to some knock on consequences that highlight what is at stake in our evaluation of its cogency. I first present a radical way of rescuing charity from the reductio above—realism about folk psychology is a tacit presupposition of the argument, and so we might reject it and rescue epistemic charity. I then move to consider ways in which the force of the argument may be wider than just epistemic charity—first by considering connections between epistemic and metaphysical principles of charity, and second by considering generalizations of the argument to other epistemic principles. Knowing what is at stake sets us up for evaluating my favoured way of responding to the reductio in section D.

C1. Rescuing charity via anti-realism about folk psychology

It is a presupposition of the argument as stated that sentences attributing intentional psychological states to people express "genuine" propositions which can be the content of further beliefs. It will be headline news if that presupposition were false. But if it were false, the argument would be resistable and epistemic charity stabilized.

Suppose that no proposition is picked out by BAD (“the psychology of this person is given by BAD”). Then the regimented argument of the previous section would have a false presupposition, since e.g. lines like “ $B(\text{BAD}|\text{E})=1$ ” would be ill-formed. What fills the first argument place of the conditional belief here, after all, needs to be a proposition.

The limit constraint and the signature of charity are also formulated in terms of beliefs with the content BAD. For example, the limit constraint tells us that when we believe/have maximal confidence in a certain proposition (e.g. about Sally’s behaviour in the control tower) we believe/have maximal confidence in a proposition about Sally’s psychological state. But in regimenting the latter this as an entailment to the conclusion $B(\text{BAD})=1$, we again suffer presupposition failure. We will need to provide some alternative regimentation of the truth behind those thoughts.

Why might one think that BAD doesn’t express a proposition? There is a long tradition in philosophy of attempting to unmask superficially declarative sentences as expressing attitudes other than ordinary belief. Moral sentences are said to express plans, rather than beliefs, conditional sentences conditional beliefs, epistemic modals states of uncertainty. On such a view, “Believing the conditional that if I flip the coin, it lands heads” does not consist in a categorical belief of the form $B([\text{if FLIP then HEADS}])$ but rather a conditional belief of the form $B(\text{HEADS}|\text{FLIP})$.⁶

Here is the crucial point. If you want to know what “degree of belief in the conditional” you have after learning some information X, you will be looking in the wrong places if you try to find a prior conditional belief in $[\text{if FLIP then HEADS}]$, conditional on X. Rather, you should look to the

⁶ The literature on no-proposition view of conditionals is large, but one starting point is (Edgington 1995).

conditional belief in FLIP given X, in HEADS given X, and in FLIP&HEADS given X. On standard Bayesian views of conditional belief, these facts would fix your conditional belief in HEADS given FLIP. And so one has a story about the evolution of the states of mind expressed by conditional sentences under the impact of new information, but it looks very different from the standard update-by-conditionalization maxim. The difference is to be expected given that “conditionals do not express propositions” and the alternative regimentation offered.

The analogous strategy for psychological interpretation would be to argue that “believing that BAD is correct” does not consist in a categorical belief of the form B(BAD), but rather in some other psychological state. What might this be? Perhaps something like this: it expresses a *plan* to treat the target of interpretation as apt to act in ways that would be rationalized by BAD. (This will do as a stand-in for now—clearly someone who wanted to seriously develop an expressivist theory of belief would have to spend a lot of time refining the proposal).

Suppose, for some genuine proposition E, rationality requires that conditionally on believing E, we must plan to treat the subject as apt to act in ways that would be rationalized by BAD. And suppose it is rationally permissible, given uncertainty about whether E is the case, to adopt a plan which is *incompatible* with treating the subject as apt to act in ways that would be rationalized by BAD (e.g. by treating them as apt to act in ways that would be rationalized by GOOD).⁷ Then we have analogues of the limit constraint and the signature of charity, but these need not be regimented in the way that lends itself to the argument of the last section.

⁷ The appeal to plans is designed to invoke the expressivism about normative concepts developed in (Gibbard 2003).

So we have a sketched a way to save epistemic charity despite the validity of the argument, and its soundness on a face-value interpretation—at a price. Realists about psychological states (including myself) will balk at the cost. But it allows us to present the upshot of the argument in a more nuanced way: rather than banning epistemic charity entirely, it only bans *realists* from endorsing epistemic charity. There is a possible argument for anti-realism here, that starts from the premise of the normative good order of charitable interpretative practices, and runs the arguments as a reductio of realism.

Such an argument has already played out in other literatures. In the literature on indicative conditionals, the expressivist, “no conditional proposition” or “no truth value” position is one of the central positions. A central role in the development of this position was played by so-called triviality arguments, developed first by David Lewis (op cit) and extended by others. Triviality arguments start from a normative premise: that rational credence in a conditional should match the corresponding conditional credence. The triviality argument reduces this assumption to absurdity, under the presupposition that conditionals express conditional propositions. Many, viewing this argument and holding fast to the normative premise, have been led to deny the presupposition by adopting the expressivist thesis about indicative conditionals. This is analogous to the line of argument for antirealism about psychological states just floated. Further, this is in fact more than an analogy. The argument that I have given against epistemic charity *is* the triviality argument, stripped down and purified, and then reinterpreted.⁸

⁸ The no-proposition view is one way expressivism about a subject matter (in a general sense) might play out, and gives one escape from the argument as stated. As a referee notes, expressivists won’t evade the simpler argument from section A—at least if they solve their Frege-Geach problems in ways that validate modus tollens (not all do). That raises another question: part of a good solution to Frege-Geach problems is handling embeddings under belief operators, and in principle this might be done that forces them to accept the premises of the argument from section B. It’s at this point in the analogous dialectic for expressivists about epistemic modals that (in my view) denials of closure under conditionalization become a natural next move. The overarching point here is that a move to expressivism/anti-realism about third-personal psychological ascriptions opens up a familiar space of moves in response to problems epistemic charity faces.

C2. Against Metaphysical charity.

The argument directly targets epistemic charity. But another group of theorists, those who endorse *metaphysical* charity, may end up caught in cross-fire.

Metaphysicians of (mental) representation are searching for a specification of the facts in virtue of which a person believes such-and-such, or desires so-and-so. Interpretationists build their account, *inter alia*, on a principle of charity. For example, (Williams 2020) I endorse and develop a metaphysics of representation based on a thesis of the following form:

What it is for x to believe that p is for the interpretation of x that makes them most reason-responsive, given their actions and in light of their evidence, to attribute to x a belief that p.

Given the total facts about x's behaviour/actions and sensation/perceptions, what is recognizably a principle of charity kicks in to determine what interpretation of the agent is correct. Various interpretations of x may be logically consistent with their actions/evidence, but it is the most favourable in a specific respect that gets things right. Metaphysical charity is charity appearing in constitutive principles such as this.

Metaphysical charity does not *entail* epistemic charity. So if epistemic charity is false, there's no direct inconsistency with continuing to endorse metaphysical charity. However, the underlying motivation for this kind of metaphysical of representation (one might insist, and the friends of metaphysical charity may agree) comes from endorsing two principles. The first would be a familiar kind of ideal-interpreter metaphysics:

What it is for x to believe that p is for an ideal interpreter of x, given full information about x's physical states and actions/evidence, to believe that x believes that p.

The second principle would be a principle of epistemic charity, which says that any interpreter *should* so far as possible attribute to agents psychological states that make them more reason-responsive. Since the ideal interpreter *does do* what any interpreter *should do*, we can derive the metaphysical charity principle from this combination. On this reconstruction, an argument against epistemic charity would undercut the rationale for metaphysical charity.

C3. Other default-reasonable beliefs.

The argument against epistemic charity exploits the *form* of the principle of charity, not its content. That means we can run the same argument against other epistemic principles that have a similar signature.

The signature of charity was the permissibility of belief beyond the evidence in conditions of uncertainty—per the main argument, this turned out to be rationally incompatible with the obligation to respect the evidence when it clearly comes into view. A willingness to respect evidence when it clearly comes into view (the limit constraint) is something that we would expect to find across the board, so in casting around for other instances of this argument we should look for other cases where (one might be tempted to think) that forming beliefs that go beyond the clearly available evidence is okay.

Some philosophers have argued that there is a class of default-reasonable beliefs—beliefs that we are permitted (if not obliged) to hold even in the absence of evidence. Examples of propositions that it is default-reasonable to believe true might include anti-sceptical assumptions such as: that an

external world exists; that this external world exhibits global regularities; and that the creatures populating this world which display apparently-intelligent behaviour have minds. If we had clear evidence against one of these anti-sceptical propositions (e.g. we consistently observe the basic regularities we assumed to be universal breaking down when we leave our home environment) then we should go with the evidence and believe them false. But absent such evidence, we continue to believe (compare e.g. Wright 2004).

The critical question concerns what happens when one is *uncertain* whether or not there is compelling evidence against one of the anti-sceptical propositions—for example, one hears reports from a somewhat unreliable source that local regularities break down (the reader should feel free to vary the track-record of the source until we get to a position where uncertainty about the existence of this evidence is appropriate). Does anti-sceptical epistemology entail that in such conditions of uncertainty it is still permissible to believe in the external world/global regularities/other minds under conditions of such uncertainty? If the answer is *yes*, then in our regimented argument we can reinterpret BAD as a sceptical proposition, E as the proposition that would be compelling evidence for BAD, if known, and run the argument just as before.

D. Charity and fragmentation.

If the argument I gave above is sound, then epistemic charity is false *if* it is construed as including what I've been calling the signature. Contraposing, if epistemic charity is to be saved, then epistemic charity can't require the combination of attitudes described in the signature. But the signature was the obvious and natural way to explain how the charitable interpreter tempers their optimism with realism. This defines a new project for the (realist) friend of charity: to say how charitable interpreters can acknowledge and reflect the evidential symmetry to which they respond, without walking into the problems we have discussed.

Let's examine what living with denial of the signature would involve. Applied to the case of Sally in the control tower, it goes as follows: the CCTV footage is glitchy enough so that *on the basis of that footage alone* (plus auxiliary inductive evidence about Sally's past behaviour, etc), what is called for is uncertainty or suspension of judgement about how Sally behaved that night. But *after* charity has been applied, and we believe that Sally was well-motivated, we reconsider the question of how she must have behaved. On the basis of our belief that she was well motivated, we form the belief that she conscientiously checked the flight path while in the control room. And since that's what we predict, we infer that this is what the glitchy footage must show. The expected direction of explanation has reversed: rather than gathering information about her behaviour in order to determine her psychological state, we retrofit our beliefs about what's going on with the footage, so that it fits with our optimistic take on her psychology.

The position just sketched is clear on the ultimate judgements that the charitable interpreter is committed to: belief in GOOD, belief in not-E, and correspondingly disbelief in BAD and disbelief in E. The same thing must be said for *every* proposition X such that X would, if learned as full information, rationally require belief in BAD.⁹

As expected, though, this "uncircumscribed optimism" version of charity leaves the charitable interpreter with a set of beliefs that don't reflect the evidential symmetry of the original situation. The friend of epistemic charity can't simply leave things like this—something about the charitable interpreter's psychology must reflect the evidential symmetry, if charitable interpreters are not to be idiotically closed-minded. So something must be added to the description.

⁹ I'm grateful for audiences in Tel Hai and Austin Texas for pressing me to say something about this sort of response to the puzzles I've introduced.

Here is my proposal. The friend of epistemic charity should posit a fragmentation in the charitable agent's response to a case like Sally's. She says that relative to one parameter-setting, the case calls for the following psychological response:

(1) uncertainty whether E, uncertainty whether BAD

but relative to some other parameter-setting, the case calls for the following psychological response:

(2) disbelief in E, disbelief in BAD.

(1) captures the idea of an epistemic symmetry in the situation, and (2) captures the result of breaking the symmetry in favour of charity. The two overall states (1) and (2) in isolation from one another are coherent and immune from the arguments of previous sections. Obviously, they conflict with each other. The original problematic signature of charity arose, according to this diagnosis, by trying to combine them into a single psychological state, with the salient attitude to E drawn from (1), and the salient attitude to BAD drawn from (2). That superposed state is incoherent, as the earlier argument showed. But if we can tell a story about epistemic charity that does justice to the sense in which (1) and (2) are both "called for", and how their separation can be maintained, then we have the elements of a satisfying diagnosis what charity consists in and how prompts, but is not sunk by, the puzzle explored in this paper.

The simplest account of a sense in which (1) and (2) are both called for would be this: the charitable agent should satisfy them both. The cost of this is obvious: charity so construed calls for agents to having inconsistent attitudes to one and the same proposition—both to be uncertain about BAD, and disbelieve it. Is this progress?

There are precedents for this sort of position. Lewis and Stalnaker, independently of the current cluster of issues, argue that agents are generally “fragmented”, with different fragments of their mind coherent on considered in isolation, but representing the world in incompatible ways (Stalnaker 1984, Lewis 1986).¹⁰ Fragmentation, in this tradition, is typically presented as a flaw—a reflection of a lack of epistemic integration due to cognitive limitations. For example, one fragment of a person’s mind might contain a belief about the area of a patch of land presented in terms of that patch’s square-meterage, another fragment an incompatible belief about the same patch presented in terms of its length and breadth. These fragments are then understood as yoked to different practical activities—verbal (questions about perimeter vs. area) and non-verbal (purchases of grass seed vs. fencing).¹¹ In the usual Lewis-Stalnaker view, under conditions of ideal a priori reflection the information in the different fragments would be pooled, and the fragments of the mental state of the agent unified.

The appeal to fragmentation by a friend of charity would have a related but distinct character. The “charitable” fragment (2) would be yoked to practical and emotional interactions with others—one would relate to Sally as is appropriate to someone of innocent character. Fragment (1) might have no (synchronic) practical manifestation, but would have a diachronic epistemic role. Specifically, upon receipt of new relevant information, fragment (1) would be updated in the first instance, and fragment (2) will be updated indirectly based on the new character of fragment (1).

A simple model of the relation between these two fragments is the following: fragment (1) contains an uncertain doxastic state represented by a set of probability functions that the agent (in that fragment) is “open to”, including some that give high probability to favourable motivations and

¹⁰ Many others under the influence of Lewis and Stalnaker have accepted such an approach. See in particular the development of these ideas in (Rayo 2013).

¹¹ The example is drawn from Rayo op cit.

some that give high probability to bad motivations. Fragment (2) is a more opinionated “charitable projection” of this, consisting of the single probability function within the set of probabilities of fragment (1) that represents the target of charity most favourably.¹²

More generally, fragment (1) would be the fragment of the charitable agent’s mind where she tracks the evolving underlying evidence, setting aside any “charitable bias”, and then we would think of charity as a mapping of uncertain epistemic states to more opinionated epistemic states under the constraint that they be “as charitable as possible”—fragment (2) being the result of charity applied to fragment (1).

Unlike the Lewis-Stalnaker rationale for fragmentation, this kind of fragmentation is not simply a product of cognitive limitations, nor is there any motivation to say that “pure reflection” would unify the fragments, or that unification is an ideal to which we should aspire. Instead, this would be an epistemology where (at least in conditions of uncertainty) fragmentation is mandatory.

This is my favourite response to the main argument against epistemic charity. I like its straightforward, thoroughly doxastic fragmentation picture. But for those friends of charity for whom this is too radical, here are two closely related variations on the proposal.

First option: so far as *beliefs* are concerned, the charitable agent should satisfy (1) and not (2). But they should satisfy an *analogue* of (2) that concerns not belief and doxastic uncertainty, but other propositional attitudes: *what is accepted for practical purposes*. On this telling, charitable agents do not genuinely believe the best of Sally, though they “accept” and so *act interpersonally as if she is innocent* until decisive evidence of guilt arises.¹³

¹² For the sets of probability model of uncertainty, see the excellent (Halpern 2003).

¹³ See (Stalnaker 1984) for a suitable notion of acceptance-states.

This picture is not really a vindication of *epistemic* charity at all. It is what one might call *pragmatic* charity—we don’t believe the best of each other, though we (should) act as if we do. So to adopt this line is strictly speaking to give up on the original ambition to vindicate epistemic charity.

Second option: so far as beliefs are concerned, the agent should satisfy (2), alone—and they should act out these beliefs in interpersonal situations straightforwardly. An analogue of (1) should be satisfied with respect to another kind of proposition attitude—a certain kind of suppositional state we might call *what-one-believes-absent-charity*. Beliefs-setting-aside-charity would have an important epistemic role in fixing belief (a bit like conditional beliefs do on a Bayesian account). What we should believe, all things considered, depends on what we believe-setting-aside-charity, plus the dictates of charity.

Here is an analogy that might make what is being suggested above less alien.¹⁴ On a simple model, the epistemology of theory choice implements a two-step process: one first eliminates any empirically inadequate theories, and among the empirically adequate theories that remain, one believes the simplest. In order to track the reasons for belief here, one will need to somehow track the interim (partially sharpened) state that occurs after step one—what theories are empirically adequate or beliefworthy-setting-aside-simplicity. The proposal above is exactly the attitudes which satisfy (1) are suppositional states of this kind. It’s worth noting, too, that if the epistemology of theory choice takes this two step form, one would expect new empirical information to impact one’s beliefs about theories *via* an update to one’s beliefs-absent -simplicity. The analogy therefore supports that the general idea present in the original fragmentation account that the attitudes satisfying (1) play a distinctive epistemic role.

¹⁴ I thank the GROUNDS research group, and Haixin Dang in particular, for suggesting this analogy.

Conclusion.

The first half of this paper developed an argument against epistemic charity—in effect, a reinterpreted version of the Lewis triviality arguments that have long been influential in other debates. Independently of what you think about the extensions and proposed resolutions presented in the second half of this paper, the puzzle is of intrinsic interest. For those who *do* wish to endorse epistemic charity (or cognate default-reasonable epistemic principles) while avoiding the most radical ways of undercutting the argument, I’ve described what I think is the most promising way forward: an epistemically required fragmentation of our belief states about others into evidence-tracking and the action-shaping components.¹⁵

J Robert G Williams

University of Leeds

¹⁵ I’ve discussed this puzzle with many people over the years, including audiences from the University of Texas, Austin, Leeds University, and Tel Hai University, and I’m very grateful for feedback at those events. The Group Thinking: New Foundations project leading to this application has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement no. 818633). For the purpose of open access, the author has applied a CC BY licence to any Author Accepted Manuscript version arising.

Bibliography

- Davidson, D. (2001) *Inquiries Into Truth and Interpretation: Philosophical Essays Volume 2*. Oxford: Clarendon Press.
- Edgington, D. (1995) 'On conditionals', *Mind*, 104/414: 235–29.
- Dretske, F. I. (1970) 'Epistemic operators', *Journal of Philosophy*, 67/24: 1007–21.
- Gibbard, A. (2003) *Thinking How to Live*. Cambridge, MA: Harvard University Press.
- Grandy, R. (1973) 'Reference, meaning, and belief', *Journal of Philosophy*, 70/14: 439–52.
- Halpern, J. Y. (2003) *Reasoning About Uncertainty*. Cambridge, MA: MIT Press.
- Lewis, D. K. (1974) 'Radical interpretation', *Synthese*, 27: 331–44.
- Lewis, D. (1976) 'Probabilities of conditionals and conditional probabilities', *Philosophical Review*, 85/3: 297–315.
- Lewis, D. K. (1986) *On the Plurality of Worlds*. Oxford: Blackwell.
- Rayo, A. (2013) *The Construction of Logical Space*. Oxford: Oxford University Press.
- Santorio, P. and Williams, J. R. G. (2022) 'Triviality and indeterminacy', *Australasian Journal of Philosophy*, 100/4: 727–42.
- Stalnaker, R. (1984) *Inquiry*. Cambridge, MA: MIT Press.
- Stroud, S. (2006) 'Epistemic partiality in friendship', *Ethics*, 116/3: 498–524.
- Williams, J. R. G. (2020) *The Metaphysics of Representation*. Oxford: Oxford University Press.
- Williams, J. R. G. (2018) 'Normative reference magnets', *Philosophical Review*, 127/1: 41–71.
- Williamson, T. (2000) *Knowledge and its Limits*. Oxford: Oxford University Press.
- Wright, C. (2004) 'Warrant for nothing (and foundations for free)?', *Aristotelian Society Supplementary Volume*, 78/1: 167–212.