Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
doi: 10.1111/phpr.12425

Policy Externalism*

DANIEL DRUCKER
University of Michigan, Ann Arbor

I develop and argue for a kind of externalism about certain kinds of non-doxastic attitudes that I call policy externalism. Policy externalism about a given type of attitude is the view that all the reasonable policies for having attitudes of that type will not involve the agent's beliefs that some relevant conditions obtain. My defense primarily involves attitudes like hatred, regret, and admiration, and has two parts: a direct deductive argument and an indirect linguistic argument, an inference to the best explanation of some strange ways we use certain conditionals. The main thought throughout is that attitudes we reason with, like belief, are very different from attitudes we don't reason with, in a way that constrains the former but not the latter. Finally, I investigate some consequences of policy externalism, including that it secures the possibility of genuine conditional apologies.

1. Introduction

Externalists of various stripes believe that a person can have a positive normative status while believing nothing about what undergirds that status. For example, a process reliabilist might think that a person can justifiably believe something simply in virtue of their belief's coming from a reliable belief-forming mechanism, regardless of whether the agent *takes* the beliefs to have been formed that way. A related kind of externalist thinks that a subject's policy to have mental states of some relevant kind in given circumstances might be highly reasonable, even though the subject has no access to whether those circumstances obtain. Such an externalist might think that the uniquely reasonable policy to have regarding what to admire is this: admire *X* just in case *X is* admirable. *Policy externalism* about *A*-type attitudes is the view that reasonable policies to have *A*-type attitudes do not essentially involve the given subject's beliefs or credences, in a way I'll make precise. I'll argue for policy externalism for a wide range of non-doxastic attitudes, including regret, admiration, resentment, and hatred; my arguments will not apply to belief itself, though, for reasons I will discuss in section 2.

I will present two kinds of argument. The first is deductive, and its conclusion is strong; the second is indirect, an abduction from our linguistic practices to a somewhat weaker conclusion. I should say, however, that even my most thorough arguments will be incomplete at best, both because of the breadth of attitudes I will discuss, and because of the ambition of the claims. My aim is not to demonstrate the truth of

* Thanks to the Graduate Student Working Group at Michigan, Sara Aronowitz, Kevin Blackwell, Paul Boswell, Anna Edmonds, Zoë Johnson King, and especially Sarah Moss and an anonymous referee for very helpful comments and discussion. Special thanks are due in particular to Eric Swanson, who provided essential advice and encouragement at every stage.

policy externalism for the attitudes I suggest, but rather to put it on the table, and to illustrate the kinds of considerations that I think should push us in its direction. Like other kinds of externalism, the view I develop conflicts with certain kinds of transparency principle, specifically that a rational and decently introspective agent will generally know what the objects of their attitudes are, barring odd Freudian cases. According to policy externalism, agents who follow the most reasonable policies concerning the relevant attitudes will very often, perhaps even typically, not know what they bear those attitudes to.[1]

2. The Direct Argument

A *policy*, as I'll use the term, is a general and standing intention to conform to a rule or set of rules.[2] As a first pass, a *rule* will be any function from circumstances to the mental states that are permissible in those circumstances according to that rule. I'll typically write out policies as imperatives ('believe p if ϕ'!), though as I'll argue in a moment, that doesn't quite exhibit enough structure for my purposes. Here's an example of a policy:

($\mathbf{I}_1$) Believe raven $n + 1$ is black if you have extremely high credence that ravens 1 through n are black and representative of ravenkind!

$\mathbf{I}_1$ is an *inductivist* policy: an agent who follows it is highly confident that some F is G when they have very high credence that some large representative sample of Fs have been G.

$\mathbf{I}_1$ is also what I'll call a *credence-involving* policy. To get at the contrast I have in mind, here is a *non*-credence-involving version of $\mathbf{I}_1$:

($\mathbf{I}_2$) Believe raven $n + 1$ is black if ravens 1 through n are black and representative of ravenkind!

The basic idea is that $\mathbf{I}_1$, but not $\mathbf{I}_2$, involves the agent's credences in the specification of the circumstances. Unfortunately, it won't do to define 'non-credence-involving' as 'a policy that doesn't mention the agent's credences in the circumstances', since the agent might have a policy to hate anyone with certain credences, say, including themselves, whether they have high credence that the individual has those credences. We need, then, to distinguish between parts of rules. Specifically, rules will divide into possible *circumstances* and the policy-holder's *relation* to those circumstances. So, $\mathbf{I}_1$ will have circumstances (raven 1 was black, etc.) and relations to those circumstances (agent has high credence that raven 1 was black, etc.). $\mathbf{I}_2$, on the other hand, will have circumstances (raven 1 was black, etc.), and relations to those circumstances (agent lives in a world in which raven 1 was black, etc.).

A policy, then, is a general and standing intention to conform to a rule or set of rules, where a rule is a function from circumstances and the policy-holder's relation to those

---

[1] The view has some affinities, then, with views that deny varieties of transparency, for example Williamson (2000)'s anti-luminosity. It more directly conflicts with the obvious generalization of what Evans (1982, p. 89) calls *Russell's Principle*, namely that "a subject cannot make a judgment about something unless he knows which object his judgment is about". (Evans finds this view in Russell (1912).)

[2] See, e.g., Bratman (1989)'s notion of a personal policy, which is close to my notion of a policy.

circumstances to the mental states that are permissible given the agent's relation to those circumstances. If our revamped $\mathbf{I}_1$ is a correct policy, for example, then an agent who has high credence that raven 1 was black, etc., ought to have high credence that raven $n + 1$ is black. On the other hand, if our revamped $\mathbf{I}_2$ is correct, if the *agent lives in a world* in which raven 1 was black, etc., they ought to have high credence that raven $n + 1$ is black. Speaking generally, it is the specified relation that does the normative work; the slot for circumstances allows for efficient bookkeeping. A policy **P** is *credence-involving* iff there is some rule in **P** that specifies that the agent must have some kind of credence or other concerning whether the circumstances of the rule obtain. Otherwise, **P** is *non-credence-involving*.

$\mathbf{I_1}$ and $\mathbf{I_2}$ are policies for having beliefs, but as I said in the introduction, my arguments have more to do with non-doxastic attitudes; doxastic ones are relevant mostly by way of contrast. So here's another credence-involving policy, a non-doxastic one, Jeffrey (1965)'s evidential decision theory's policy for extrinsic preference (roughly, instrumental preference):

(1) Prefer p to $\neg p$ just in case $\sum_{w \in W} u(w) \cdot Cr(\{w\}|p) > \sum_{w \in W} u(w) \cdot Cr(\{w\}|\neg p)$,

where $u(w)$ is the utility the agent would get from w's being actual and Cr is the agent's credence function.[3] This seems like a reasonable policy, since extrinsic preference is what makes actions rational, and the agent's information matters to what counts as rational.

Policy externalism about *A*-type attitudes is the view that all the most reasonable *A*-type policies to have are non-credence-involving.[4] In the rest of this section and the next, I'll present arguments for policy externalism for attitudes like hatred, regret, and admiration. (The exact list of attitudes is unimportant for my purposes; I'm just attempting to make the position plausible for many of them.)

Dropping anger for a modest slight when the object of your anger is contrite and swears not to do similar things in the future strikes me as a good policy. That is, it's good (fitting, etc.) to forgive in such circumstances. Regretting the hurtful things one's done also seems like a good policy. Neither policy, though, mentions *finding out*. Suppose a token action, *X*'s φ-ing, has the properties of hurting someone and being voluntary or intentional. It seems like a good policy to regret it. We don't need to add that it also has the property of *being known by X to have the first two properties*. That makes us and our mental states too central to having the attitude. When philosophers specify what makes an object worth

---

[3] This assumes the set of worlds W is finite, which I'll assume throughout.

[4] The example I've already given, to admire X when X is admirable, might give the impression that policy externalism embraces only *narrow-scope* norms for these attitudes. A narrow-scope norm has the form $\ulcorner \phi \supset O\psi \urcorner$, where '$O$' is some normative operator, whereas a wide-scope norm has the form $\ulcorner O(\phi \supset \psi) \urcorner$. The practical difference is that when a narrow-scope norm's antecedent is true, agents are required to make the consequent true; but with wide-scope norms, agents may also go on to make the antecedent false. 'Admire X when X is admirable!' is indeed best represented as a narrow-scope norm, but the policy externalist needn't join Kolodny (2005) in rejecting wide-scope norms; she is only committed to the norms' being non-credence-involving. A wide-scope norm (e.g., to not be angry with one person who did you a significant injustice, or to be angry with all such people) might be perfectly reasonable to the policy externalist. Thanks to a reviewer for pressing me to clarify this point.

bearing some attitude like love or hatred to, usually it has to do with features of the object itself, like its beauty or its maliciousness—and not our credences.[5]

Here's a case to pump your intuitions in policy externalism's favor:

> *Implausible Lack of Regret*. Lisa has lived on the moral edge for a long time; while she's always *tried* to do the right thing, she's tried to stay right on the right side of the permissible. Statistically speaking, she's incredibly likely to have in fact done something wrong—hurt someone, say. Suppose, knowing all this, she says:

(2) I've performed no action I should regret.[6]

To my ear, Lisa sounds very overconfident: she should think (2) is very likely false.

In addition to that sort of example, here's another test for seeing whether our credences figure into our policies. We have strong reason to think they don't iff it seems wrong for the strength of the attitude to co-vary with one's confidence that some relevant circumstances obtain. We have no such strong reason with extrinsic desire, since something like (1) (or a causal version) seems to be sensible enough.

Imagine a series of counterparts $X_1, \ldots, X_n$ across $w_1, \ldots, w_n$, respectively, all of whom think Y might be F. They all think F is a feature that makes someone worthy of hatred—they just have different credences in the proposition that Y is F. (Suppose, to simplify matters, that Y is F in w_1 iff Y is F in w_2 iff ... iff Y is F in w_n.) Say their credences range from .75 to .95. Should X_n hate Y *more* than X_{n-1}, who should hate Y more than X_{n-2}, and on down the line? It seems the more reasonable position is: there's some amount or range of amounts it makes sense to hate Y, and the X_is differ not in how much they should hate Y, but in how confident they should be that they should hate Y to that degree.

As confirmation of this, suppose someone's being F makes them *very very* worthy of hatred. But now suppose X is very confident that Y is *not* F; .95, let's say. It is *not* the case that X ought to hate Y some significant amount. We don't—and shouldn't—hate in proportion to the expectation to someone's worthiness of hatred. Modifying hatred in proportion to expectation seems like a mistake. In contrast, it's plausible that credences are expectations of truth-values,[7] in which case one's credence in p obviously should vary according to your confidence that some relevant circumstances obtain (e.g., the p-circumstances). On the present view, hatred, resentment, and the rest are more like *guesses*; one doesn't "guess more strongly" that p when one is more confident that p.[8] This is not to say these attitudes aren't degreed, but rather the degrees don't work in the way credences do, since they don't co-vary with expectations that way.[9]

Well, why aren't they *exactly* like guesses? Reasonable policies about them would then be minimally credence-involving, just having a bit about whether a given credence meets the required threshold to make the attitude reasonable. So, on this view, a

---

5 See, e.g., the literature on fitting attitudes, which is too massive to survey here. It starts, perhaps, with Ewing (1947); see also D'Arms and Jacobson (2000), Rabinowicz and Rønnow-Rasmussen (2004), and Zimmerman (2011), for a taste.

6 I put it this way to rule out the pattern of morally risky behavior as something deserving regret.

7 For this idea, see Jeffrey (1986).

8 Guessing still is credence-involving to some extent. See Horowitz (forthcoming) for discussion.

9 The basic thought is similar to the thought that we shouldn't *punish* in proportion to expected guilt. See, e.g., Buchak (2014); but for a contrasting view, see Rosenberg (1984), at least as applied to liability in torts.

reasonable policy for hatred might be to hate X if I have credence above T that X is worthy of hatred. The answer is that it'd be even better to hate those who are *worthy of hatred*, credences aside, if we could. If we hated just in that way, we would hate fewer people who didn't deserve it, and hate more people who did. Our hatred would be more fitting and more worthy of our endorsement. So, if it's possible to hate in line with those policies, those are the most reasonable ones to adopt; and if it's not possible for A-type attitudes, then that would be a pretty strong argument against policy externalism about A.[10]

More carefully, here's the argument, abstracting over attitude-types:

P1. All the best A-type policies to have are non-credence-involving, if we can conform to them.

P2. We can conform to some of the best non-credence-involving A-type policies.

C1. So, the best A-type policies to have are non-credence-involving.

P3. A policy **P** for A-type attitudes is unreasonable to have if there are clearly better A-type policies than **P** that one can have.

C2. So, the most reasonable A-type policies are non-credence-involving.

I've already gestured at the strategy for defending P1: namely, if we conform to non-credence-involving policies, we won't be misled into anger against the innocent, admiration of the noxious, and pride in the worthless. Similarly, we won't be misled *out of* admiration of the admirable, love of the lovable, and regret of the things worth regretting. These benefits would be great if we could secure them. Moreover, they would not be *unfitting* ways to have these attitudes. A full defense of P1 would have to look at individual attitudes and policies, of course; but I think it's clear enough that if there were no access constraints on which of these policies we could follow, the best policies for these attitudes would not be credence-involving. So, even though there's a lot more to be said about P1, I will move on, and spend much more time defending P2.

3. For Information Independence

To show that we can conform to non-credence-involving policies for attitudes like regret and hatred, I'll provide a mechanism. Then I'll address an objection to the mechanism, which will provide occasion to explain why credence-involving policies are reasonable when they are, and why they aren't when not.

The mechanism is simple and familiar, though not in this connection. Suppose X has a policy with the following form:

(3) a. Hate X just in case X has properties $F_1, \ldots, F_n$.
    b. Hate X just in case X kicked your dog.

[10] This is similar to to the argument that Wedgwood (2002) gives for policy internalism about belief (as I interpret him). For an earlier antecedent, see Pollock and Cruz (1999).

Knowledge that X has these policies will justify the following ordinary-language self-ascriptions:

(4) a. I hate whoever has properties $F_1, \ldots, F_n$.
b. I hate whoever kicked my dog.

Here's one from the wild:[11]

(5) I HATE whoever invented TV commercials.

These utterances self-ascribe *generalized attitudes*, attitudes an individual can bear to *whatever* has the necessary features. Distinguish these from *particularized attitudes*, such as hating just N, whatever their properties may be. X bears a generalized attitude A toward Y iff (i) X bears A to Y in virtue of Y's being $F_1, \ldots, F_n$, (ii) X bears A to every Z that has properties $F_1, \ldots, F_n$ in virtue of having those properties, and (iii) for every possible $Z \neq Y$, were X to remain intrinsically the same and were Z to have features $F_1, \ldots, F_n$, X would bear A to Z. A generalized attitude is an attitude one bears to different things in virtue of their having some common set of properties and where variation in the objects the subject bears the attitude to can happen without any intrinsic change in the subject.

This distinction between generalized and particularized attitudes resembles the distinction between *notional* and *relational* attitudes, and therefore also between *de dicto* and *de re*. That distinction is illustrated by the following:[12]

(6) a. Ralph believes that there are spies.
b. There is someone whom Ralph believes to be a spy.

This looks like it's easily analyzed by a scope distinction;[13] but now consider:

(7) Perseus seeks a gorgon.

This has at least two readings: the relational one, if there is a specific gorgon Perseus seeks, and the notional one, if Perseus merely seeks *some* gorgon (but none in particular). It also has no obvious place to make our earlier scope distinction. How best to capture the distinction between these two readings isn't important for my purposes; I just want to stress that my distinction is a different distinction.

A generalized attitude is like a notional one, in that there are some properties of the object the attitude hooks onto; the subject's attitude cannot miss its target in virtue of being directed at something with the wrong properties. If I bear a relational attitude to x that is motivated by my taking x (perhaps under some guise) to be $F_1, \ldots, F_n$, I might nevertheless end up hating something with none of those properties. Thinking I'm seeking a gorgon, I pursue Medea; but I've only confused her with Medusa, the real gorgon. So generalized attitudes are not relational attitudes. Generalized attitudes will also

---

11 This is the name of a Facebook group: https://www.facebook.com/I-HATE-whoever-invented-TV-commercials- 102522833124796/.

12 See Quine (1956).

13 (a) is something like '$\textit{BEL}$(Ralph, $(\exists x)(\textit{SPY}(x))$', and (b) is something like '$(\exists x)(\textit{BEL}$(Ralph, $\textit{SPY}(x)))$'.

resemble relational attitudes, though, in that they are *specific*: if Karin is the one who kicked my dog, then in virtue of subscribing to (8b), I'll hate *Karin* specifically. This isn't just because Karin is the only one who in fact kicked my dog (as we might suppose). It's that my attitude will be in part about *her*, grounded in the fact that *she* in particular kicked my dog. So generalized attitudes are not notional ones. So generalized attitudes are neither relational nor notional.

That said, the distinction between notional and relational readings of attitude reports is notoriously slippery, and so you might think that either of the differences between generalized attitudes and relational or notional attitudes that I appealed to is unreal. In particular, generalized attitudes might just be relational attitudes that cannot miss their targets. One way that might go is to say that generalized attitudes are relational ones that we have *in virtue of* having corresponding notional ones. What matters to me, though, is just this: we can have attitudes to *specific* objects that *unfailingly* hit their targets.

Why think that we can have these attitudes? Philosophical orthodoxy already provides the relevant materials. Kripke (1972) popularized the idea of a proper name whose reference is fixed by description. Take the following:

(8) Let 'Pat' denote the inventor of the idea that babies are delivered by storks. I'm grateful to Pat for a good laugh or two!

Attitudes like the one self-ascribed in (8) fit the above characterizations: the speaker's attitude targets Pat in virtue of their having invented the stork story, and had someone else done it instead, it would've targeted them.[14] In some instances we'll want not just singular but plural terms. That's no problem:

(9) Let 'The Jerks' denote the people that put my car on the building's roof. I hate The Jerks![15]

Caveat: while this is a strategy that *guarantees* that we can perform the relevant feats, I don't claim that every time we do, we use names like this.

Anyhow, here is the mechanism. Suppose that X has a policy to bear A to whichever objects $x_1, \ldots, x_n$ are such that x_1 has $F_1, \ldots, F_m, \ldots, x_n$ has $G_1, \ldots, G_k$. X can conform to this policy by employing mental names whose references are fixed as in (8) and (9), thereby *unfailingly* bearing A to $x_1, \ldots, x_n$. There is nothing especially mysterious about any of this. That's why you sometimes find people saying things like this:

(10) Hillary cares about me.[16]

---

[14] This idea comes from McKinsey (2009). For a different but similar idea, I could have instead appealed to Kaplan (1978, 1989)'s 'dthat', an expression that takes a description as argument and outputs a directly referring singular term.

[15] 'I hate The Jerks!' needs to be able to be read *distributively* (I hate this Jerk, and that Jerk, etc.) rather than collectively (I hate The Jerks as a group). See Link (1983) for an analysis of the distinction.

[16] This comes from multiple places. Here's one: http://www.washingtonpost.com/wp-dyn/content/article/2008/05/19/AR2008051902729.html.

This can be even if Hillary's never even met the speaker. We don't find it very hard to believe that people can bear attitudes like care, hatred, and the rest even to people they don't know at all.

All this leads to an objection. You might worry that if I'm right about what I've said so far then epistemology crashes. That's because this seems like a great doxastic policy:

(11) Assign credence 1 to p just in case p is true, and assign credence 0 to p just in case p is false.

If it's possible for someone to conform to this policy, then it's possible for an agent to satisfy the following self-ascription:

(12) I am fully confident of whatever propositions are true, and fully doubtful of whatever propositions are false.

If, as many think, I am supposed to believe in line with whatever credences have least expected inaccuracy, then it seems like I could adopt no better policy than (11).[17] Even if you think other values go into determining the goodness of a credence function, like fit with evidence or understanding, it should be clear that agents *do not* have credences in line with (11), not even those epistemologists who *do* think that accuracy of credences is all that is of ultimate epistemic value.

I think this is explained by the fact that generalized belief, by which I mean belief in whatever propositions satisfy a given description, is not always possible, even when one in fact endorses a given policy that would otherwise generate that generalized belief. If there's been a murder, and while Raval is the prime suspect, the evidence is equivocal, I cannot appropriately say:

(13) I believe whatever's true about whether Raval is the murderer. So, if he is, I believe that he is, and if he isn't, then I believe that he isn't.

So, generalized belief is not always possible even if one endorses the policy that would otherwise generate the belief. That makes sense given something like the following:

> Reasoning with Beliefs. If S believes that p, then S can use p to reason practically and theoretically, i.e., S can use p as a premise in the (conditionally) rational generation of a sufficiently wide range of new beliefs, assignments of subjective probabilities, preferences over acts, and other attitudes.[18]

This is a minimally functionalist account of belief. According to it, you can't be said to believe a thing unless you can use it to rationally change some of your other mental states and perhaps behavior. I think it's true and explains the difference between belief and the other attitudes, but I want to note that my argument would still go through even

[17] For the accuracy framework, see Joyce (1998, 2009), Pettigrew (2016), and Schoenfield (2015).

[18] Epistemologists have recently advanced similar requirements on belief—not just rational or justified belief, but belief itself. Ross and Schroeder (2014)'s *Reasoning Disposition* account of belief is most closely connected. For a very similar thought applied to *credences*, see Joyce (2009, p. 263).

if it isn't part of the right explanation. The present objection—that epistemology would be far too easy were what I said true—fails because examples like (13) show generalized belief isn't always or typically possible, whereas similar examples for hatred, regret, resentment, and the rest are both common and felicitous, as I'll argue in section 3. The particular explanation of that failure doesn't matter for the *truth* of policy externalism; it does help us understand why it's true, though, and so I will give and defend the explanation from Reasoning with Beliefs.

Reasoning with Beliefs points to an important problem with self-ascriptions like (13): the putative belief would not put me in a position to rationally choose my actions. Suppose that Raval is in fact the murderer. Then if I said (13) truly, I would believe that he was. But if I really believed Raval was the murderer, I should behave quite differently: I should cease getting people to suspend judgment about him, or I should at least start thinking of him as a scoundrel; and I should cut my business dealings off with him, as well. Yet in the situation as described, I would be silly to do any of those things.[19] So I can't really believe in line with (13).[20]

This explanation raises two related questions. First, what are we to say about this?

(14) Let 'Jack' denote the murderer. I believe that Jack is the murderer.[21] So, if Raval is Jack, then I believe he is; and if he isn't, then I believe he is not.

And second, why is nothing like Reasoning with Beliefs true of the attitudes I've been concerned with, like hatred or regret?

In response to the first problem, I'll note that, to get the contrast between belief and the other attitudes, and thus to answer the challenge from (11) and (12), general purely descriptive beliefs suffice:

(15) I believe whatever's true concerning whether our world is deterministic. So, if it is, I believe it is, and if it isn't, I believe it isn't.

---

[19] If you don't see the silliness, we can make the evidence equivocal between nineteen people in addition to Raval, so that my credence in Raval's guilt should only be roughly .05.

[20] This, incidentally, explains why utterances like (10) don't ascribe *conditional* attitudes. A conditional attitude is a generalization of the notion of conditional belief. A conditional belief is, roughly, a belief we have *conditional* on the truth of something else. For example, I have the belief that Biden will not win the election in 2020, conditional on his not running, but I don't now *believe* he won't win in 2020, since he might run. The worry for what I've said so far is that the generalized attitudes are really conditional ones. So, for example, according to this worry, (8) self-ascribes gratitude *conditional on* Pat's having invented the stork story, and (10) ascribes Hillary *conditionally* caring about the speaker (presumably the proposition the care is conditional on is recoverable somehow from context). There are some big problems with saying this, though. For one, the speaker in (8) *knows* that Pat invented the stork story, and so the gratitude cannot be merely conditional. (If Pat is Aesop, say, then the speaker won't know that Aesop invented the stork story, but that's not a problem if we invoke some kind of modes of presentation. I'll elaborate on that in a moment.) Second, if (13) self-ascribed a conditional belief, rather than an unconditional belief that Raval is the murderer (so long as he in fact is), then it should sound perfectly all right, since the corresponding conditional belief would be as reasonable as conditional beliefs get. (13) sounds bad, though, so it doesn't merely ascribe a conditional belief. We have reason to think that constructions with 'belief' swapped out for 'hate', etc., would work similarly, so I take it that these conditional constructions don't *just* ascribe conditional attitudes. Thanks to a reviewer for pressing this objection.

[21] Let's suppose in this scenario I have some good evidence there was a unique murderer.

(15) will be as bad as (13), and for the same reason, namely that REASONING WITH BELIEFS is true. So we have our contrast.

Nevertheless, (14) raises interesting issues. I think it must be dealt with in whatever way such cases involving names whose reference is fixed by description. It is in fact some evidence for what I've been saying that philosophers have tried so much *to* prevent examples like (14) from being true. Donnellan (1977), for example, argued that names like 'Jack' do not make *de re* beliefs about the referent possible; but of course, *if* Raval is the murderer, I (let's suppose) already can have many *de re* attitudes toward him. More helpful for present purposes is Schiffer (1977)'s *hidden-indexical* account. On this kind of a view, a sentence of the form ⌜*X* believes that *N* is *F*⌝ has the following form:

(16) BEL (*X*, ⟨*F*, *N*⟩, *m*),

where *m* is a mode of presentation of *N*, which one in particular to be decided by context. Then we can explain the relevant parts of (14), when they are acceptable, as having the following form:

(17) (BEL (me, ⟨MURDERER, Jack⟩, m_1) & Jack = Raval) → BEL (me, ⟨MURDERER, Raval⟩, m_2)

Suppose Jack is Raval. If m_1 is a murderer-y mode of presentation, but m_2 is just Raval's normal mode of presentation, then (17) won't actually be satisfiable by reasonable people with the stipulated evidence. That's because such a belief ought to be usable, by REASONING WITH BELIEFS, i.e, I ought to be reasonable in doing the things I mentioned earlier, such as cutting my business ties with Raval or whatever. That would not be reasonable in the present circumstances. If something like the hidden-indexical account is right, epistemology doesn't crash, even for *de re* predications.[22]

This kind of maneuver makes the second question more pressing. For example, we need an explanation of (18)'s felicity:

(18) If Karin kicked my dog, then I hate her!

If modes of presentation somehow prevent examples like (15) from sounding felicitous, why do they not block (18) from sounding felicitous? The answer is that nothing like REASONING WITH BELIEFS holds of hatred. Why not, though?

The answer is that we don't *reason with* attitudes like resentment, love, or the rest. We do reason with beliefs, i.e., with what we take to be true. By contrast, even if I regret hurting you, I reason with the (assumed) fact that I regret hurting you, or that I hurt you, *not* with the regret itself. In other words, I reason with my beliefs about the matter.[23] It is then no mark against our having a particular resentment against *X* that

[22] I don't insist on Schiffer's account in particular. Related ones such as Crimmins and Perry (1989) would do as well.

[23] This point originates, as far as I can tell, with Stampe (1987), who makes it about desire.

we cannot use it in our reasoning, since we don't use any resentment in our reasoning, at least not directly.[24]

The other attitudes have different functions. Hating *X*, for instance, makes it prudent to avoid *X*; it is no *less* prudent to avoid *X* when one hates *X* but doesn't know that one does. Knowing that one hates *X* might make it more *reasonable* to avoid *X*, but reasonableness and prudence are different things.[25] It is bad to do something you hate doing, even if you don't know you do or should hate doing that thing. Similarly, loving *X* makes it more prudent to interact in the right ways with *X* and to treat *X* especially well, even if it doesn't make doing so more reasonable. More generally, these non-doxastic attitudes are *orientational*: they make certain things, courses of action, etc., good or bad, or prudent or imprudent for us, whether or not we know this. We can evaluate given non-doxastic attitudes for reasonableness; but unlike beliefs, their primary normative contribution concerns not the reasonableness of further beliefs, actions, etc., but their prudence, or their goodness.

They're not *just* orientational, of course. Very often they're also motivational. But what they motivate need not be very sensitive to ways of describing: anger might motivate revenge against the person who kicked my dog, however else I think of them to myself; whereas if I *believed* Karin was the dog-kicker, I'd be able to reason to all sorts of other things (she was disingenuous when she pretended to like my dog). None of the anger roles—or, though this is just speculation before a more detailed study, admiration roles, regret roles, etc.—seem to place any constraints on information, since they don't play a role in reasoning.

So, modes of presentation will not be obstacles to bearing a particular attitude like hatred, since modes of presentation prevent certain kinds of *reasoning*, not the orientational phenomena I just discussed. Modes of presentation affect only those attitudes that contribute directly to reasonableness. So, P2 is true of, or at least plausible of, policies for many non-doxastic attitudes. C1 then follows.

---

[24] What about desires? You might think that we reason with them, which would cause trouble for my explanation because the following seems felicitous:

(i) If you would be happiest being a lawyer, then I want you to be a lawyer, and if you would be happiest being a doctor, then I want you to be a doctor.

That would mean we can have generalized attitudes even with attitudes we reason with. Personally, I do think (i) is perfectly fine when, say, said by a mother to her daughter. But I doubt we reason with desires directly. It can be paralyzing when we're ignorant of what we want. The simplest explanation of that is that we need beliefs about our desires for them to affect our reasoning. Now, you might think that decision theory is a formalization of ordinary reasoning with desires. I think of it somewhat differently: it sets a (subjective) standard, using our actual credences and utilities, against which the wisdom of various courses of action can be measured and compared. To apply the theory—that is, to use it in practical reasoning—we need to have beliefs about what those utilities are, a highly nontrivial task. On this picture of decision theory, we need never use it to reason with our desires directly. Broome (2013, p. 268) shares this skepticism about reasoning with desires, but does think we can reason with preferences. But in Broome (2006), where the issue receives fuller discussion, he's inclined to think that "a preference may be nothing other than a belief about goodness" (pp. 207–8). For my own part, I think that we don't even reason with preferences, rather than with either beliefs about goodness or beliefs about the preferences themselves, for the reasons I just gave. The arguments in this footnote are unfortunately only preliminary; I intend to discuss these matters in much greater detail elsewhere. Thanks to a reviewer for pushing me to address these worries.

[25] There's a use of 'prudent' inherited from Aristotle's '*phronesis*' that means something like 'practically wise', and thus looks more like reasonableness. I mean 'prudent' in roughly the sense of 'good choice for the agent's welfare'. Thus I use it in roughly Bricker (1980)'s sense.

What about P3? Here it is again:

P3. A policy **P** for *A*-type attitudes is unreasonable to have if there are clearly better *A*-type policies than **P** that one can have.

This is an instance of a general norm not to pick outranked options. If P3 is false, it will be because of the general category of the supererogatory. This category has proved troublesome to integrate into our overall deontic scheme, and moreover it's not clear that there is an *all-things-considered* supererogatory. I remain attracted to P3, personally, but I am also comfortable with a slightly weaker claim, one that looks more like the first externalism I mentioned:

C2′. Some of the most reasonable policies for *A*-type attitudes are non-credence-involving.

While I do believe the stronger and more ambitious claim, even C2′ has striking consequences.

That concludes the direct argument. In the next section, I'll give a different argument, and in the process use policy externalism to explain some otherwise puzzling data.

4. The Linguistic Argument

4.1. Introducing the Data

Attitudes and conditionals are two perennial sources of philosophical puzzles. Here's one that arises from their interaction.

Begin with the following cases:

> *Implausible Regret*. On walking back from a chat with Jof, Jöns wonders whether he had culpably offended him. He says to himself:

(19) If I hurt Jof's feelings, I seriously regret doing so.

Unbeknownst to Jöns, he *did* hurt Jof's feelings.

> *Implausible Hatred*. Mia was watching Jof and Jöns, and saw what very well might have been Jöns culpably offending Jof. She thinks to herself:

(20) If Jöns hurt Jof's feelings, I resent him for it.

Again, unbeknownst to Mia, Jöns *did* hurt Jof's feelings.

> *Implausible Forgiveness*. Jof's feelings were hurt by Jöns. But he says to himself:

(21) If (but only if) Jöns seriously regrets offending me, I forgive him.

These conditionals are perfectly ordinary language—watch closely and you might catch yourself saying one from time to time. But given (19)–(21) and the facts, we can infer that Jöns seriously regrets hurting Jof's feelings, that Mia resents Jöns for hurting Jof's

feelings, and that Jof forgives Jöns for hurting his feelings. That's bizarre, because Jöns doesn't even believe that he hurt Jof's feelings; neither does Mia; and Jof doesn't believe that Jöns seriously regrets doing so—he doesn't even know whether she believes she offended him in the first place. For those reasons it can seem false that Jöns has those regrets, that Mia carries that resentment, and that Jof forgives Jöns. Yet the only inference rule we used was *modus ponens*. So the *Implausible* cases have counter-intuitive consequences if they generalize. First, a person can have these attitudes without knowing that the antecedents obtain. In light of that it can be hard to see how (19)–(21) can be reasonable things to say. Another is that we're massively ignorant of the objects of regret, resentment, forgiveness, hatred, admiration, desire, and more. We've learned to live with similar conclusions, for example that our grasp of what we say is often very incomplete.[26] But my cases have nothing to do with the vagaries of content determination. Finally, Jöns, Jof, and Mia seem to have none of these attitudes' typical phenomenologies.

Since (19)–(21) are so ordinary, we should be reluctant to conclude that they cannot be true, known,[27] or reasonable in the relevant circumstances. That leaves us with the following options. First, we can *reinterpret* them so that we cannot detach them in the relevant circumstances; we can deny the unrestricted validity of *modus ponens*; or we can take (19)–(21) at face value *and* accept that we can detach them even when the speakers don't know the antecedents obtain. I, of course, think we should take this third option. Most of the work to do this was in fact already done in sections 1 and 2. So, first, I'll briefly explain how that kind of policy externalism can help explain the *Implausible* cases. Then, since I intend for this section to constitute an inference to the best explanation, I'll show how the other strategies for accounting for those cases don't work.

Policy externalism's explanation—or rather, any explanation that turns on something like C2′—is very simple. Take (19). Jöns has a policy to regret the things he's done that hurt other people's feelings, say. The conditional expresses that policy, and it can be true even if Jöns doesn't, even can't, know that he hurt Jof's feelings by the same mechanism I discussed in section 2. Because (19) can be true, and because it's *reasonable* to have a policy of regretting hurting people's feelings (culpably and needlessly, say), Jöns's utterance can be perfectly reasonable. We can give the same kind of explanation of (20) and (21), too. The upshot is that all the speakers come out sincere and reasonable.

The other two strategies for dealing with (19)–(21) have severe but informative problems.

4.2. Against Reinterpretation

There are a couple of different reinterpretation strategies to try. First, one can try out a scope distinction.

(22) Tantalus ought to serve someone their children in a stew if he wants to exact on that person the most terrible sort of revenge.

It shouldn't follow that Tantalus ought to serve his children in a stew if he does in fact want to exact the most terrible sort of revenge. A standard way to avoid this problem is to say that (22) really has (23)'s logical form (with '*O*' for 'ought'):

---

[26] See, e.g., Putnam (1975) and Burge (1979).

[27] If they weren't known, they would run afoul of the knowledge norm of assertion (see, e.g., Unger (1975) and Williamson (2000).

(23) $O(\phi \supset \psi)$.

(23) isn't in the right form for *modus ponens*.[28] Perhaps we can think of (19)–(21) as involving wide-scoping.

Unfortunately, that strategy won't work. First, that kind of strategy only works if the attitude in the consequent has a clausal complement. This is possible with some attitudes, but it seems that it's not possible with all, e.g., 'resents' in (20). In other words, we have reason to think that not all intentional attitudes ultimately take only propositions as objects.[29] And even for those that do, there are some problems. So, the propositional paraphrase of (19) is:

(19′) I seriously regret that, if I hurt Jof's feelings, I did so.

In other words, this has Jöns regret a tautology. That is obviously a terrible paraphrase of (19), but it seems the best that can be done to get the wide-scope strategy going. So, I think the wide-scope strategy won't work.[30]

An initially plausible idea is to treat (19) as elliptical for something like the following:

(19^{+}) If I hurt Jof's feelings and find out, I will seriously regret doing so.[31]

The combination of discovery in the antecedent and future tense in the consequent makes these conditionals much less worrisome than (19)–(21).

This strategy also won't work. The first problem is relatively superficial: it cannot handle deathbed cases. So, suppose Karin, knowing Death has come at last for her, says:

(24) If Antonius kept all his vows to me during his long time abroad, I'm grateful.

In fact, Antonius did keep all his vows (let's suppose Karin doesn't, even can't know that). So the simple version of this strategy won't work. It's not for lack of sophistication that it fails, but for a more general reason. Instead of (19^{+}), suppose we interpret (19) as:

(19^{++}) If I hurt Jof's feelings and I were to find out, I would seriously regret having done so.

This does avoid the deathbed problem. Nevertheless, "Thomason" conditionals like (25) suggest a different problem:[32]

(25) If Sally's deceiving me, I'll never know it.

Now consider the following:

---

28 Greenspan (1975) is the *locus classicus* of the approach, and the source for (22).

29 See, e.g., Forbes (2000) and Montague (2007).

30 For an unrelated battery of arguments against wide-scoping strategies in the case of *modus ponens* failures for natural-language 'ought', see Silk (2014).

31 See von Fintel (2012, p. 29) for this idea applied to the *Miners' Puzzle* (see below).

32 So-called because van Fraassen (1980) attributes them to Richmond Thomason.

(26) But even if Antonius broke some of his vows, if he has taken pains to hide that, I'm grateful that I'll at least never know that he did break his vows.

To reinterpret (26) along the lines of (19^{++}), we would get:

(26^{++}) Even if Antonius broke some of his vows, if he has taken pains to hide that fact but I found out, I would be grateful that I would at least never know that he broke his vows.

This has or entails something with the form $\ulcorner(\phi \wedge \psi) \boxdot\!\!\rightarrow (\chi \wedge \neg\psi)\urcorner$, which can only be true on most semantics if $\ulcorner\phi \wedge \psi\urcorner$ is impossible.[33] So, this strategy fails because of Thomason conditionals: there are felicitous conditionals with the problematic consequences that cannot be elliptical as proposed.

Finally, you might think these are "biscuit" conditionals, named for Austin (1970)'s example:

(27) There are biscuits on the sideboard if you want any.

I don't think this line will help. First, (19) takes 'then', unlike biscuit conditionals:[34]

(19^{*}) If I hurt Jof's feelings, then I seriously regret doing so.

So I have some doubts that these are biscuit conditionals. But beyond that, a biscuit conditional $\ulcorner\psi$ if $\phi\urcorner$ seems to entail ψ, e.g., that there *are* biscuits on the sideboard. That's exactly the entailment that the reinterpretive strategies were aimed at avoiding.

Reinterpreting (19)–(21) doesn't seem promising. So, if we want to find another way to deny the relevant commitments, we have to try something else.

4.3. Against Information Dependence

(19)–(21) are not the only examples that lead to problematic commitments, given *modus ponens*.[35] I'll focus on a case that has drawn a lot of attention, the *Miners' Puzzle*:[36]

> *Miners' Puzzle*. Ten miners are trapped and all of them are either in A or B. Block can block A, block B, or do nothing. If he blocks A, they all live if they're in A and all die if they're in B; and if he blocks B, they all live if they're in B and all die if they're in A. If he does nothing, one person dies and the other nine survive.

Typical judgments:

(28) a. Block ought not to block A and ought not to block B.
b. If the miners are in A, Block ought to block A.
c. If the miners are in B, Block ought to block B.

---

33 See, e.g., Stalnaker (1968) and Lewis (1973).

34 See Davison (1979) and Iatridou (1994).

35 See also McGee (1985).

36 The case originates with Regan (1980), and it received prominent discussion in Parfit (1984).

(28a) is intuitive because blocking either shaft is very risky given Block's information; the expected utility is much lower than blocking neither shaft. The trouble is that (28b–c) entail the negation of (28a), at least if we use the stipulation that the miners are all in A or all in B, as well as classical logic, specifically.[37]

One solution is to reject *modus ponens*.[38] Here's the idea. Let i, i', etc. range over *information states*, sets of worlds capturing the information of some relevant party. Say that i *accepts* ϕ at t iff, for all $w \in i$, ϕ is true at w and i at t.

(29) $[\![\ulcorner \text{if } \phi, \psi \urcorner]\!]^c = 1$ iff ψ is accepted at the time of the context at every $i' \subseteq i$ such that ϕ is accepted at the time of the context at i' such that there is no $i'' \supset i'$ such that ϕ is accepted at the time of the context at i''.[39]

Next, a *selection function* as a function from information states i to the set of worlds considered optimal by the lights of that function.[40] A deontic selection function, e.g., will probably only select worlds where all promises made have been kept. Then, where i_g is the contextually relevant information state, $\ulcorner \text{ought } \phi \urcorner$ is true if every world that's deontically best given the information state is a world in which ϕ is true. More precisely:

(30) $[\![\ulcorner \text{ought } \phi \urcorner]\!]^c = 1$ iff $(\forall w)(w \in d(i_g) \supset w \in [\![\phi]\!]^c)$,

So long as which worlds the deontic selection function picks can be altered by the information-state updating procedure in (29), then we can avoid the worrying commitments. Since (28)'s speaker is stipulated *not* to know (or have beliefs about, etc.) the antecedents in (b) or (c), we are not compelled to accept that the speaker is committed to whichever consequent corresponds to the actually true antecedent.

To adapt this solution to the attitude verbs in (19)–(21), we need to give them lexical entries that can make use of the indicative conditional's ability to shift the information state. To see how this might be done, we should look at 'want'. To start, notice that we can construct a *Miners' Puzzle* for desire:

> *Miners' Puzzle for Desire*. The miners are trapped as before, and Block accepts the typical judgments, i.e., that he ought to do nothing, but that if they're in A, he ought to block A, and if they're in B, he ought to block B. So he says the following:

(31) a. If they're in A, then I want to block A;
b. and if they're in B, I want to block B.
c. But, I don't want to block either of them, since I ought not to block either of them.

Given that the miners are either all in A or all in B, (30a–b) entails:

(32) Either I want to block A or I want to block B.

---

37 See Kolodny and MacFarlane (2010) for the simple derivation.

38 See Kolodny and MacFarlane (2010) and MacFarlane (2014, ch. 11) for further details.

39 See MacFarlane (2014, p. 270).

40 I'm making Lewis (1973)'s LIMIT ASSUMPTION, i.e., that there is always a set of deontically optimal worlds. Not much here hangs on it.

This contradicts (31c), but even if it didn't, (32) is still odd: why would the speaker be ignorant of which shaft they want to block? This isn't the typical Freudian case of repressed desire—the conditionals suggest that which of A or B Block wants to block somehow depends on which shaft the miners are in.[41]

We can mimic the solution just described for 'ought', but we need to enrich our information states i with an algebra over the set of worlds in i closed under complementation and union and with a probability measure Pr over that algebra.[42] Then we say that i accepts a non-probabilistic sentence ϕ at t just in case for all $w \in i$, ϕ is true at w and i at t, and i accepts an at least partly probabilistic ϕ just in case ϕ is true at all worlds in w evaluated with Pr. For example, consider the following:

(33) The probability of its raining in Chicago on April 2, 2020 is greater than .5.

i accepts (33) just in case Pr (<it rains in Chicago on April 2, 2020>) >.5. Or suppose we have the following conjunction:

(34) The probability of its raining in Chicago on April 2, 2020 is greater than .5 and dogs bark.

i accepts (34) just in case all $w \in i$ are worlds in which dogs bark, and Pr(<it rains in Chicago on April 2, 2020>) >.5 according to i's Pr. Finally, we need to update our conditional semantics to reflect our new information states:

(35) $[\![\ulcorner \text{if } \phi, \psi \urcorner]\!]^c = 1$ iff ψ is accepted at the time of the context at every $i' \subseteq i$ such that (i) ϕ is accepted at the time of the context at i', (ii) if there are $\neg\phi$-worlds in i, i''s probability measure Pr^{ϕ} is i's probability measure Pr conditionalized on ϕ[43] (otherwise i''s probability measure is Pr), and (iii) there is no $i'' \supset i'$ such that ϕ is accepted at the time of the context at i''.

With all that said, here's Levinson (2003)'s semantics for 'want'.[44] Let u_S be S's utility function at the relevant context, and Pr be the relevant information state's probability measure. Then S wants p to be true just when p's expected value (by S's and i's lights) is higher than $\neg p$'s. In symbols:

(36) $[\![\ulcorner S \text{ wants } \phi \urcorner]\!]^c = 1$ iff $\sum_{w \in W} u_S(w) \cdot Pr(\{w\}|p) > \sum_{w \in W} u_S(w) \cdot Pr(\{w\}|\neg p)$,

This gets the desired result: the conditionals (31a, b) are true, but the negation of (c) doesn't follow.[45]

---

41 Of course, this is totally expected on the theory presented in sections 1 and 2, if it also applies to desire.

42 For this general strategy, Moss (2013), Swanson (2016), and Yalcin (2012).

43 Pr^{ϕ} is Pr conditionalized on ϕ just in case, for all x and ψ such that $Pr(\psi|\phi) = x$, $Pr^{\phi}(\psi) = x$.

44 This lexical entry goes well with Weirich (1980)'s view about conditional desire sentences $\ulcorner$if ϕ, $\psi\urcorner$, namely that they express high utility in ϕ on the indicative supposition that ϕ. See also Charlow (2013).

45 (36) might seem to take sides between evidential and causal decision theory (see Joyce (1999) for an opinionated introduction to the controversy). You might further worry that no semantics for 'want' should encode *any* decision theory, even the correct one. Since I'm only exploring, not proposing the entry in (36), I can accept that objection. See Carr (2015) for this worry applied to 'ought'.

The first thing I'd like to point out about (36) is that I haven't put any restrictions on who *i*, and so *Pr*, can be tagged to. For the *Miners' Puzzle*, this seems right: (28a–c) are third-personal. But things get odder with the version involving 'want'. Consider the following:

(37) a. If they're in A, then Block wants to block A;
b. and if they're in B, Block wants to block B.
c. But Block doesn't want to block either of them, since he doesn't know which shaft the miners are in.

If I try, I think I can hear versions of (36a, b) that sound all right in the stipulated circumstances. These readings call to mind examples like Williams (1981)'s gin case: if the man thinks the liquid on the table is gin, but it's really gasoline, we can say the following to him:

(38) You don't want to drink that![46]

But if I get myself to hear (37a, b) this way, (c) sounds bad. More importantly, though, lexical entries like (36) won't actually solve our problem. Return to (19): if we were somehow able to rig up a lexical entry for 'regret' that makes its truth depend on the relevant information state, we third parties should be able to reason as follows:

(39) a. If Jöns hurt Jof's feelings, he seriously regrets doing so.
b. He *did* hurt Jof's feelings.
c. So, he seriously regrets hurting Jof's feelings.

We can know Jöns hurt Jof's feelings without Jöns knowing, bringing back our problematic consequence.

We might, then, say that *Pr* in (36) and its potential analogues is somehow restricted to the *speaker's* information state, so that the utterances in (28) come out true but not in (37) and (39). In other words, perhaps the reasoning in (31) *only* works first-personally. Rather than work out an implementation of this idea, I'll point out a problem that would affect *any* implementation: it's looking more and more like the truth-conditions of (19) will turn out to be uncomfortably close to (19^{+})'s, or its most sophisticated versions. On this view, what matters for detachment isn't that *some* relevant information state is updated with the antecedent, but *Jöns's*. But we've already seen that this reinterpretation strategy fails because of examples like (26) (repeated here):

(26) But even if Antonius broke some of his vows, if he has taken pains to hide that, I'm grateful that I'll at least never know that he did break his vows.

If the strategy under discussion were right, then (26) should sound awful—pointless, because it'd be *impossible* to detach. Yet it sounds perfectly ordinary. So, I think Thomason conditionals give us in-principle reasons to reject any strategy like the ones I've been discussing in this section.

[46] For interesting discussion of this example, see Korsgaard (2008).

Another problem with (36) is that it only captures *extrinsic* desire. Yet there are (19)-like examples with intrinsic desire:[47]

(40) If pleasure is good for its own sake, then I want everyone to have as much pleasure as they can.

(36) cannot make good sense of (40). Indeed, nothing *like* (36) can, since certain kinds of intrinsic desire—utilities over entire worlds—are provably rationally unchanged by updates to probability functions.[48] So imagine someone says:

(41) If utilitarianism is true, I want this world to be the best it can be as far as utilitarianism is concerned.

No information-updating strategy can capture this sentence.

Finally, (31a, b) have so-called "non-reflecting" readings. For example, consider:

(42) If they're in A, I still want to block neither, since I don't know they're in A.

The original *Miners' Puzzle* has similar readings:[49]

(43) If they're in A, Block still ought to block neither, since he doesn't know they're in A.

The trouble is that (36) cannot capture the non-reflecting readings like the one brought out by (42), even if we make (36) more sophisticated by restricting *Pr* to the agent's credences. It's interesting whether (19) has a non-reflecting reading. Consider:

(44) If I hurt Jof's feelings, I don't regret doing so, since I don't know that I did.

This sounds callous, at least to my ear. Policy externalism about regret, the idea that reasonable policies for regret are non-credence-involving, can help to explain this: just like *Implausible Lack of Regret*, (44) expresses a credence-involving policy—a bad one, even an immoral one, namely *not* to regret what one doesn't know was wrong. Someone who thinks they *might* have done a particular something worth regretting should be in a

---

47 For a helpful discussion of the distinction between intrinsic and extrinsic desire, see Arpaly and Schroeder (2014).

48 Let $v_X: \mathcal{P}(W) \rightarrow \mathbb{R}$ be a function that records how desirable X finds prospects, satisfying the following axioms:

Normality. $v_X(W)=0$.

Averaging. $v_X(p \vee q) = \frac{v_X(p)Cr_X(p) + v_X(q)Cr_X(q)}{v_X(p) + v_X(q)}$,

with $Cr_X(p) = \sum_{w_i \in p} Cr(\{w_i\})$ and $v_X(p) = \sum_i v_X(\{w_i\})Cr_X(\{w_i\}|p)$. Then we can define how desirable X finds p conditional on q as follows: $v_X(p|q) := v_X(p \wedge q) - v(X)$. Suppose $v_X(\{w\}|p) > v_X(\{w\prime\}|p)$ and $Cr_X(\{w\}), Cr_X(\{w'\}) > 0$. Then $v_X(\{w\}) - v_X(p) > v_X(\{w'\}) - v_X(p)$, so that $v_X(\{w\}) > v_X(\{w'\})$, since w and w' are both p-worlds. This reasoning is reversible. So, an agent's ranking of worlds by conditional subjective desirability cannot come apart from her ranking of worlds by unconditional subjective desirability. See Bradley (2009) for further details and discussion.

49 See, e.g., Cariani et al. (2013).

different state of mind than someone who doesn't think they might have. (44) sounds bad because we can implicitly see this.

Any unambiguous attitude-expression *V* of whose corresponding attitude-type policy externalism is true will not generate genuine *Miners' Puzzle*-like cases, since the (c)-lines will be false. And for the information-dependence strategies to work, the (c)-lines would have to be true. When *Miners' Puzzle*-like cases genuinely pull in two ways and where ambiguity is absent, they compel us to have a consistent policy, be it credence-involving or non-credence-involving.

The *Miners' Puzzle for Desire* is one such case. First, let's rule out the existence of a relevant ambiguity as best we can. Some ambiguities have been claimed, but they wouldn't license the differences between (a, b) and (c).[50] Nor would the distinction between intrinsic and extrinsic desires help, since they are all instrumental.

The crucial question is what to say about this:

(45) If they're in A and I don't know that they are, I want to block neither.

Because conditionals don't obey antecedent strengthening,[51] none of (31a–c) entails an answer to (45). (As we saw with Thomason conditionals in section 2, updating on an antecedent is very different from updating on the proposition that one *knows* the antecedent.) Moreover, it seems to me that (36) gets odd results here. It says that (45) is true *only if* the utility of saving all ten miners' lives when you don't know that they're in A is less than the utility of saving just nine miners' lives when you don't know they're in A. Some might find this palatable, but I doubt most who want to affirm (45) will.[52] So, there doesn't seem to be ambiguity present, and it seems that a uniform interpretation of 'want' leads to bad results. As far as extrinsic desire goes, there are at least *prima facie* reasonable credence-involving and non-credence-involving policies. Which is of these to pick is, it seems to me, a very substantive matter, one on which reasonable people can disagree. That seems not to be true of regret, hatred, and the rest. And if (36) seems to require unreasonable utilities to obtain results that people would be reasonable to accept, then (36) isn't plausible as a semantics for 'want'.

Finally, we're in a position to understand why (36a–c) are hard to hear as collectively acceptable in a single context, even though we *can* hear them as individually acceptable in different contexts. The policies they assume Block has are individually reasonable, but they cannot all be reasonably held *together*, since they conflict.

Even were (36) true, it would not extend to the other attitudes in conditionals like (19)–(21); we've also seen reason to think that (36) is false. For these reasons, the information-dependence strategies won't work. Since the reinterpretive and information-dependence strategies fail, we should accept the account in sections 1 and 2. In the next section, I'll briefly explore some of the consequences of that account.

---

[50] Davis (1984) distinguishes between volitive and appetitive desires—roughly between desires that lead to action and desires that register what a person would like. Levinson (2003) makes the same distinction, using 'motivational' for 'volitive' and 'partial' for 'appetitive'. Lewis (1988) makes a similar distinction between cool and warm desire. The trouble is that all the desires in *Miners' Puzzle for Desire* either do or could fall on the volitive side of this line.

[51] Antecedent strengthening is the schema $\ulcorner$if $\phi, \psi\urcorner \vDash \ulcorner$if $\phi \wedge \chi, \psi\urcorner$.

[52] von Fintel (2012), following Kratzer, embraces this consequence, at least for 'ought'.

5. Lessons and Consequences

I've given arguments that the following two conclusions are true of many attitude-types:

C2′. Some of the most reasonable *A*-type policies are non-credence-involving.

C2. The most reasonable *A*-type policies are non-credence-involving.

C2 entails C2′. The direct argument I sketched in section 1 aimed to establish C2, but the linguistic argument in section 3 at best only establishes C2′. It's worth drawing out some of their consequences, and some consequences of how I defended them.

Since Kaplan (1968) and Kripke (1972), there has been haggling over what the consequences of reference-fixing by description are for belief. It seemed to give us bizarre powers to learn contingent things from the armchair just by coming up with the right names. Where learning is not an issue, though, similar consequences don't seem so bizarre. Those are the consequences I used to make sense of the conditionals (19)–(21). The fact that not all of our attitudes involve the manipulation of information leads directly to the fact I've argued for, that some expected limitations don't exist. This has some real practical consequences. I'll illustrate one, conditional apologies.

Begin with a case very close to the cases with which I began. Imagine that Jöns says the following:

(46) If I hurt your feelings, I'm very sorry.

This is a *conditional* apology. These might seem dubious, and often they should. We're all familiar with paradigmatically insincere examples. But suppose Jöns's ignorance of the antecedent is non-culpable: Jof has not approached him to extract an apology, and the evidence in Jöns's possession really was equivocal. Then whatever we say about whether (46) expresses a genuine apology, it will at the very least not fail to do so for being insincere or rude in the way those paradigmatic examples are.

Well, *can* (46) express a genuine apology? I think that it can.[53] There are at least two reasons to think that it can't: that Jöns doesn't have the belief that what he did was wrong (in the particular way it was), and that he can't actually *feel* sorry.[54] The belief condition is, I think, misguided, exactly because it rules out cases where the apologizer's evidence is genuinely equivocal.[55] More importantly for present purposes, if what I've argued is correct, Jöns *can* feel genuinely sorry for what he's done, even if he doesn't have the belief that what he's done is wrong, and even if he doesn't exhibit typical remorseful behavior, or experience typical remorseful phenomenology. If this is so, we need not say that remorse isn't really required for an apology to be genuine.[56]

---

[53] *Pace* most of the literature on the subject, with the notable exception of Miller (2014). Our reasons differ, though.

[54] Bovens (2008) thinks both are necessary for an apology to be genuine.

[55] Miller makes this point.

[56] Miller changes the remorse requirement, for example, to a *conditional* remorse requirement. If I'm right, that's unnecessary and unmotivated.

Policy externalism allows us to see how we can apologize even in situations of limited information. Similarly, we can *forgive* in such cases, too: Jof's (21) (repeated here) feels like a natural thing to say.

(21) If (but only if) Jöns seriously regrets offending me, I forgive him.

Policy externalism allows us to engage in certain social activities that would otherwise be impossible, since these activities depend in part on our having certain attitudes that would otherwise be either impossible or unreasonable.

I want to close by discussing a thorny set of issues. I argued we can have these generalized attitudes, and it's worth investigating that more thoroughly. The policies I had in mind were things like:

(47) Admire *X* just in case *X* is admirable!

Now, I argued we could satisfy them, in the sense that there is no difficulty in principle of our doing so. But it is no secret that we *do* admire all sorts of people who are not admirable, for example if they appear so. We know we do by the typical phenomenology. Generalized attitudes like hating whoever kicked your dog *feel different* from the particularized attitude of hating Peter, who seems to you like he kicked your dog. It's not as though I feel *nothing* when I hate whoever kicked my dog, or regret whatever I've done that's hurt people; but it does often feel less intense or visceral.

The overall situation is a bit puzzling. If we won't go astray in our attitudes if we just stick to the obviously good policies, that is, if we only ever have the obviously fitting generalized attitudes without the perhaps misled particularized attitudes, then why should we have the particularized attitudes as often as we do? That question strikes me as similar to the question of why people use (standard) proper names rather than descriptions or proper names whose reference was fixed by description. For one, it can be less cognitively taxing to do so. In our assertions and thoughts we sometimes sacrifice guaranteed accuracy for cognitive efficiency, especially where we're highly confident that the proper name and whatever description you'd use to be safe co-refer. Other times we're just inattentive to the difference, since, for example, when we're convinced that a particular person did something horrible, it's hard to feel the need to distinguish between the two ways we can hate them. Even if the best policies to have are clear, we can be lax about following them in situations in which either way seems to get us to the same place. Finally, just like doxastic attitudes, emotions are not completely (or probably even mostly) voluntary. A policy to weigh evidence rationally and disinterestedly can be hard to stick to when doing so threatens our self-conception; and a policy not to have particularized attitudes can be hard to follow, even if clearly best, when it'd feel good to hate a particular person *and know it*.

This is where worries about the phenomenology involved in our attitudes return. As I said, I do feel *something* when I regret the things I've done that have hurt people unnecessarily. It is true, though, that such feelings are likely to be far less visceral than the regret I'll feel when I know I hurt *this specific person* when I did *this specific thing*. The phenomenology's intensity and unpleasantness makes the regret feel more sincere, since it seems like we're hurting ourselves to make amends in a way that regrets like those expressed in (19) and (46) don't. That doesn't make those attitudes any less real, or

motivating. If my regret is genuine, I should investigate thoroughly and tread lightly in similar situations. An attitude isn't realer just for corresponding to or causing a more intense or visceral experience. Our inability to stick to generalized attitudes seems to me to derive in part from placing too much importance on phenomenology.

References

Arpaly, Nomy. and Schroeder, Timothy. 2014. *In Praise of Desire*. Oxford, UK: Oxford.

Austin, J. L. 1970. "Ifs and Cans." In *Philosophical Papers*, 205–232. Oxford, UK: Oxford. 2nd ed.

Bovens, Luc. 2008. "Apologies." *Proceedings of the Aristotelian Society* 108:219–239.

Bradley, Richard. 2009. "Becker's Thesis and Three Models of Preference Change." *Politics, Philosophy, and Economics* 8:223–242.

Bratman, Michael E. 1989. "Intentions and Personal Policies." *Philosophical Perspectives* 3:443–469.

Bricker, Phillip. 1980. "Prudence." *Journal of Philosophy* 77:381–401.

Broome, John. 2006. "Reasoning with Preferences?" In Serena Olsaretti (ed.), *Preferences and Well-Being*, 183–208. Cambridge, UK: Cambridge.

—— 2013. *Rationality Through Reasoning*. Malden, MA: Wiley-Blackwell.

Buchak, Lara. 2014. "Belief, Credence, and Norms." *Philosophical Studies* 169:285–311.

Burge, Tyler. 1979. "Individualism and the Mental." In Peter French, Theodore E. Uehling, Jr, and Howard K. Wettstein (eds.), *Midwest Studies in Philosophy IV: Studies in Metaphysics*, 73–121. Minneapolis, MN: University of Minnesota Press.

Cariani, Fabrizio. Kaufmann, Magdalena. and Kaufmann, Stefan. 2013. "Deliberative Modality under Epistemic Uncertainty." *Linguistics and Philosophy* 36:225–259.

Carr, Jennifer. 2015. "Subjective Ought." *Ergo* 2:678–710.

Charlow, Nate. 2013. "Conditional Preferences and Practical Conditionals." *Linguistics and Philosophy* 36:463–511.

Crimmins, Mark and Perry, John. 1989. "The Prince and the Phone Booth: Reporting Puzzling Beliefs." *Journal of Philosophy* 86:685–711.

D'Arms, Justin and Jacobson, Daniel. 2000. "The Moralistic Fallacy." *Philosophy and Phe- nomenological Research* 61:65–90.

Davis, Wayne A. 1984. "The Two Senses of Desire." *Philosophical Studies* 45:181–195.

Davison, Alice. 1979. "On the Semantics of Speech Acts." *Journal of Pragmatics* 3:413–429.

Donnellan, Keith. 1977. "The Contingent A Priori and Rigid Designators." In Peter French, Theodore E. Uehling, Jr, and Howard K. Wettstein (eds.), *Midwest Studies in Philosophy II: Studies in the Philosophy of Language*, 12–27. Minneapolis, MN: University of Minnesota Press.

Evans, Gareth. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Ewing, A. C. 1947. *The Definition of Good*. London, UK: Routledge and Kegan Paul.

Forbes, Graeme. 2000. "Objectual Attitudes." *Linguistics and Philosophy* 23:141–183.

Greenspan, P. S. 1975. "Conditional Oughts and Hypothetical Imperatives." *Journal of Philosophy* 72:259–276.

Horowitz, Sophie. forthcoming. "Accuracy and Educated Guesses." In Tamar Gendler and John Hawthornee (eds.), *Oxford Studies in Epistemology*, volume 6. Oxford: Oxford.

Iatridou, Sabine. 1994. “On the Contribution of Conditional Then.” *Natural Language Semantics* 2:171–199.
Jeffrey, Richard. 1965. *The Logic of Decision*. New York: McGraw-Hill.
—— 1986. “Probabilism and Induction.” *Topoi* 5:51–58.
Joyce, James M. 1998. “A Nonpragmatic Vindication of Probabilism.” *Philosophy of Science* 65:575–603.
——1999. *The Foundations of Causal Decision Theory*. Cambridge, UK: Cambridge.
—— 2009. “Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief.” In Franz Huber and Christoph Schmidt-Petri (eds.), *Degrees of Belief*, 263–297. Springer.
Kaplan, David. 1968. “Quantifying In.” *Synthese* 19:178–214.
—— 1978. “Dthat.” In Peter Cole (ed.), *Syntax and Semantics*, 221–243. Academic Press.
—— 1989. “Demonstratives.” In Joseph Almog, John Perry, and Howard Wettstein (eds.), *Themes from Kaplan*, 481–563. New York: Oxford University Press.
Kolodny, Niko. 2005. “Why Be Rational?” *Mind* 114:509–563.
—— and MacFarlane, John. 2010. “Ifs and Oughts.” *Journal of Philosophy* 107:115–143.
Korsgaard, Christine M. 2008. “The Myth of Egoism.” In *The Constitution of Agency*, 69–99. Oxford, UK: Oxford.
Kripke, Saul. 1972. “Naming and Necessity.” In Donald Davidson and Gilbert Harman (eds.), *Semantics of Natural Language*, 253–355, 763–9. Dordrecht: D. Reidel. Revised edition published in 1980 as *Naming and Necessity* (Harvard University Press, Cambridge, MA).
Levinson, Dmitry. 2003. “A Probabalistic Model-Theoretic Semantics for Want.” In R. Youngand Y. Zhou (eds.), *SALT XIII*, 222–239.
Lewis, David. 1973. *Counterfactuals*. Oxford: Blackwell.
—— 1988. “Desire as Belief.” *Mind* 97:323–332.
Link, Godehard. 1983. “The Logical Analysis of Plurals and Mass Terms: a Lattice-Theoretic Approach.” In R. Bäuerle, C. Schwarze, and A. von Stechow (eds.), *Meaning, Use, and the Interpretation of Language*. Berlin: de Gruyter.
MacFarlane, John. 2014. *Assessment Sensitivity*. Oxford: Oxford.
McGee, Vann. 1985. “A Counterexample to Modus Ponens.” *Journal of Philosophy* 82:462–471.
McKinsey, Michael. 2009. “Thought by Description.” *Philosophy and Phenomenological Re- search* 78:83–102.
Miller, Kristie. 2014. “Conditional and Prospective Apologies.” *Journal of Value Inquiry* 48:403–417.
Montague, Michelle. 2007. “Against Propositionalism.” *Noûs* 41:503–518.
Moss, Sarah. 2013. “Epistemology Formalized.” *Philosophical Review* 122:1–43.
Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.
Pettigrew, Richard. 2016. *Accuracy and the Laws of Credence*. Oxford, UK: Oxford.
Pollock, John and Cruz, Joseph 1999. *Contemporary Theories of Knowledge*. Towota, NJ: Rowman & Littlefield.
Putnam, Hilary. 1975. “The Meaning of Meaning.” In Keith Gunderson (ed.), *Language, Mind and Knowledge, volume 7 of Minnesota Studies in the Philosophy of Science*, 131–93. Minneapolis: University of Minnesota Press.

Quine, W. V. O. 1956. “Quantifiers and Propositional Attitudes.” *Journal of Philosophy* 53:177–87.

Rabinowicz, Wlodek and Rønnow-Rasmussen, Toni. 2004. “The Strike of the Demon.” *Ethics* 114:391–423.

Regan, Donald. 1980. *Utilitarianism and Cooperation*. Oxford: Oxford.

Rosenberg, David. 1984. “The Causal Connection in Mass Exposure Cases: A “Public Law” Vision of the Tort System.” *Harvard Law Review* 97.

Ross, Jacob and Schroeder, Mark. 2014. “Belief, Credence, and Pragmatic Encroachment.” *Philosophy and Phenomenological Research* 88:259–288.

Russell, Bertrand. 1912. *The Problems of Philosophy*. London: Williams and Norgate. Paper-back edition by Oxford University Press, 1959.

Schiffer, Stephen. 1977. “Naming and Knowing.” *Midwest Studies in Philosophy* 2:28–41.

Schoenfield, Miriam. 2015. “Bridging Rationality and Accuracy.” *Journal of Philosophy* 112:633–657.

Silk, Alex. 2014. “Why ‘Ought’ Detaches: Or, Why You Ought to Get with My Friends (If You Want to Be My Lover).” *Philosophers’ Imprint* 14:1–16.

Stalnaker, Robert C.1968. “A Theory of Conditionals.” In *Studies in Logical Theory: American Philosophical Quarterly Monograph Series, No. 2*. Oxford: Blackwell.

Stampe Dennis W. 1987. “The Authority of Desire.” *Philosophical Review* 96:335–381.

Swanson, Eric. 2016. “The Application of Constraint Semantics to the Language of Subjective Uncertainty.” *Journal of Philosophical Logic* 45:121–146.

Unger, Peter. 1975. *Ignorance*. Oxford: Oxford University Press.

van Fraassen, Bas. 1980. “Review of Brian Ellis, Rational Belief Systems.” *Canadian Journal of Philosophy* 10:193–197.

von Fintel, Kai. 2012. “*The Best We Can (Expect) to Get?*” Paper for a session on Deontic Modals at the Central APA, February 17, 2012.

Wedgwood, Ralph. 2002. “Internalism Explained.” *Philosophy and Phenomenological Research* 65:349–369.

Weirich, Paul. 1980. “Conditional Utility and Its Place in Decision Theory.” *Journal of Philos- ophy* 77:702–715.

Williams, Bernard. 1981. “Internal and External Reasons.” In *Moral Luck*, 101–113. Cambridge: Cambridge.

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

Yalcin, Seth. 2012. “Context Probabilism.” In Maria Aloni (ed.), *Logic, Language, and Meaning*, 12–21. Heidelberg: Springer.

Zimmerman, Michael J. 2011. “Partiality and Intrinsic Value.” *Mind* 120:447–483.