Attitudes Beyond Belief: A Theory of Non-Doxastic Attitude Formation and Evaluation

by

Daniel Drucker

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2017

Doctoral Committee:

Associate Professor Eric Swanson, Chair
Associate Professor Sarah Moss
Marshall M. Weinberg Weinberg Professor Brian Weatherson
Associate Professor Ezra Keshet

2017

Acknowledgments

I'll be brief; there's no way to fully express my gratitude here. So, thank you to Eric Swanson for being the best possible committee chair, conscientious, empathetic, and wise; to Sarah Moss, for the intense and challenging conversations that always left me wanting to be and do better; to Brian Weatherson, for showing the importance of a view of the whole of philosophy; and to Ezra Keshet, for his generosity and example(s).

Thanks to Michigan's (sometimes honorary) philosophers, who've been my friends and family for six years, including: Chloe, Sara, Boris, Dave, Gordon, Kevin, Mara, Paul, Bryan, Francesca, Sarah B., Victor, Lingxi, Mercy, Guus, Alex, Matt, Josh, Sydney, Zoë, Jim, Maria, Ishani, Eduardo, Neil, Filipa, Adam, Cat, Alvaro, Chip, Steve, Umer, Patrick, Nils, Rohan, Jamie, Rich, Damian, Nina, and Elise. It's incredible that this kind of community exists; I'm unspeakably lucky to have been a part of it. The proof of this is how impossible it feels to leave it.

Thanks to Plato and Agnes Callard, who shaped my worldview in ways I still don't fully understand.

Thanks to Hakeem Jerome Jefferson, one of the only non-philosophers on the list, for teaching me as much as any philosopher could.

Thanks to Dmitri Gallow, one of the best friends I could ever have, whether ten feet away or a thousand miles away. You're one of the sharpest people I know, and one of the kindest.

Thanks to Anna Edmonds. I have too many things to thank you for to thank you for any one thing. I can't believe I get to be your friend.

And thanks to my mother, Cathy Drucker. Your strength is hard to fathom, and I'll never underestimate it. Thank you for believing in me. I'll always try to earn it.

TABLE OF CONTENTS

ABSTRACT

Attitudes Beyond Belief: A Theory of Non-Doxastic Attitude Formation and Evaluation

by

Daniel Drucker

Chair: Eric Swanson

I present and explore a normative theory of non-doxastic attitudes like desire, hatred, and admiration. The viewpoint is general and abstract: independent of any particular flavor or source of normativity, I explore general features any acceptable way of forming these attitudes would have, especially in contrast to doxastic attitudes like belief. The first three chapters present a relatively unified picture of non-doxastic attitude formation, grounded in types of non-doxastic attitudes we can have in contrast to their impossible doxastic analogues. In particular, I defend a kind of externalism; I apply it to a specific attitude, desire, to show how the approach works; and then based on a different aspect of the view, I present what I take to be some of the best ways to motivate the view, and some of the best ways to criticize it, as well. In the last chapter, I develop a theory about how these attitudes are best communicated, this time meant to capture how belief is best communicated, too, though it is inspired by certain features of non-doxastic attitudes.

CHAPTER 1

Introduction

Belief plays an enormously large role in philosophy. We have an entire subfield, epistemology, devoted to which beliefs are good in which ways, how to get those good beliefs, and how to transmit them. This is all important and edifying work. There are many more attitudes than belief, though, like desire, regret, admiration, and hatred. I fear that too much focus on belief, as opposed to these other attitudes, distorts our work on these attitudes, and causes us to miss important generalizations that apply to *every* attitude. The goal of my dissertation is to begin to remedy this situation. I do so by presenting a normative theory of attitude formation, applying it to one of the most important non-doxastic attitudes, desire, and then presenting a normative theory of attitude expression. I say 'begin' because there's still a good deal more work on this specific theory to be done. In each chapter, I try to be explicit that I think that even if I have succeeded at everything I've intended to do, I will only have established what I've presented as plausible options. Further work would go beyond mere plausibility. It would also have more applications to specific attitudes than just desire. I hope to do that work soon, and I also hope that others may join me in trying.

Before I get into the specifics of the chapters here, I want to say a little about my basic methodological and philosophical orientation. You will notice some of these chapters have many numbered example sentences, even though they ultimately aim to do more than argue for linguistic theses, however interesting and important those are in their own right. There are two reasons I appeal to what seems fine appropriate, true, and reasonable to say so frequently. First, linguistic considerations often reflect pre-theoretic common sense. Often when something seems implausible because of some of our theoretic commitments, it can be useful, both epistemically and dialectically, to see that in other contexts we *assume* the thing that seemed so implausible. Of course, sometimes theory wins out, and sometimes it *should* win out. When I appeal to linguistic considerations, I also try to give a theory as to why speaking that way would and does reflect the truth. Those theories do not render the appeals to linguistic intuition and use otiose, though, because the theories and the data are mutually reinforcing: we should trust the data more when we have a theory that makes

sense of them, and we should trust the theory more if it comports with how we naturally think. Some may think that common sense, especially relatively untutored common sense, shouldn't be of much importance in philosophy, and I cannot answer them persuasively here. I'll just note my disagreement.

The second reason I appeal to linguistic considerations is Davidsonian, which is that how we speak (rather trivially) determines what phenomena we're talking about. There simply could not be massive mismatch between our subject-matter and the way we talk about that subject-matter, because then otherwise we'd be talking about some other subject-matter. What sorts of error count as changing the subject and what counts as, well, simple error—false belief—is a fraught question. Again, I won't attempt to say anything that will convince anyone about where to draw the border here. I'll just note that when we commonly and unhesitatingly speak in a certain way, we should think that it likely does fix the phenomena of interest.

The other thing I would like to emphasize is this: when I say I am giving a normative theory of the attitudes, I mean 'normative', not 'moral'. One of the more *a priori* surprising sociological facts about contemporary philosophy is that normative theories of belief are very typically not principally concerned with morality's requirements on belief, if there are any, but normative theories of the other attitudes are usually heavily focused on the moral and the ethical. (I cannot prove this, say with statistics. I could very well be wrong. But it is my strong impression.) Belief is assessed from the perspective of rationality, often qualified as specifically *epistemic* rationality. Desire and preference are a more mixed case—there are interesting discussions of both moral and rational requirements on desire. Nevertheless I do think the asymmetry holds for the other attitudes. A good deal of the initial impetus for this dissertation was a reaction to realizing this. It seems to me there are interesting things to say about the rationality of non-doxastic attitudes, even the non-doxastic ones besides desire, and that we are in the very early days of saying those things.

When we turn to them, though, as I said at the beginning, we need to be very careful that we don't simply mimic or transport principles or considerations from belief that might not apply to the other attitudes. A heuristic I've employed throughout the dissertation is to find the ways in which given classes of attitudes are different from other classes in ways that affect the large-scale structural generalizations that are true of them. Much of the dissertation focuses on one of those differences, the possibility of general attitudes. But it isn't the only one, and I suspect there are others, which means I suspect that will continue to be a fruitful heuristic going beyond this dissertation.

In what follows, I will briefly summarize the contents of each chapter, so that the reader can pick whichever of them seems most interesting to them. Every chapter is motivated by

the concerns I just mentioned, and more or less employs some of the methodology I just described, but it is not a continuous work. Chapters 2 through 4 all concern the same basic theory, but they do so from very different vantage points. In Chapter 2, I argue for the importance and reality of general attitudes, and how it motivates a kind of externalism. In Chapter 3, I apply the basic outlook from Chapter 2 to the specific case of desire, and show what happens when we do so. In Chapter 4, I use similar tools from Chapters 2 and 3 to argue for a different position, namely that, essentially, for most attitudes we should form particular (that is, non-general) attitudes only by way of forming general attitudes. Chapter 4 also addresses what I took in writing the dissertation to be the most pressing objections, ones I am still wrestling with myself. Each of the three chapters is readable on their own, even though they all relate to one another in complex ways. Chapter 5, which is about communication, is of course quite different from these. Nevertheless, I figured it was important to include because of the way in which it coheres with the general motivations behind the dissertation, namely to stop focusing so much on belief (and, in communication, assertion). It is accordingly even more self-contained than the other chapters, but I hope it doesn't appear too out of place.

That said, here are the summaries. In Chapter 2, "Policy Externalism", I present and motivate a variety of externalism that I call policy externalism: that, for many attitudes besides belief and probably intention, the best formation policies to have will not involve beliefs in any central way. An example of a policy like this would be to admire X just when X is admirable; a rival would have it that we should admire X just when you *believe* X is admirable. I try to show how we can actually conform to those policies by introducing the notion of a general (in that paper, 'generalized') attitude.[1] I then argue that this doesn't generate problematic consequences for epistemology. I then show how we're committed to something like policy externalism by the way we speak, in particular some strange-seeming but completely common conditionals. Much of that discussion is devoted to showing that rival interpretations, which are anyway linguistically unnatural, cannot account for all of the data. Also included is a discussion of the *Miners' Puzzle* for these attitudes, and a discussion of how conditional apologies can be genuine apologies, even if the speaker doesn't even believe they did anything wrong.

In Chapter 3, "Wanting What One Should Want", I argue that we can have specific general desires, in particular that we can want what we objectively should want, all things considered. This possibility, perhaps surprisingly, generates a subjective requirement, and thus I think a rational requirement, to have certain desires *because* they are the most moral

[1] I kept the terminology because the paper has since been published, and I didn't want the dissertation version to differ from the published version. I apologize if this is confusing in any way!

ones to have. I then compare it to rival approaches to rational desire, including decision theories, and argue that it seems to come out ahead, at least in some ways. I then deal with three objections: that the approach would lead to unlucky incoherence, that the difference between it and its rivals is really non-substantive, and that we really can't have the desires I say we can have.

In Chapter 4, "Deliberation, General and Particular", I introduce a view I call deliberative generalism, roughly the view that in deliberating about which attitudes to have, we rationally should only consider general features of things that make having those attitudes to those things worthwhile, rather than the things themselves. I give a very elementary argument for it, and along the way motivate a different principle that seems to me intuitively correct, essentially that if we bear a given attitude to something for a given reason, we should bear that attitude to anything to which the reason applies. I also address a number of objections, the most interesting of which (to my mind) is that generalism would lead to agents' being alienated from their attitudes.

Finally, in Chapter 5, I introduce the notion of a Φ-assertion (roughly, an attempt at getting one's addressee to have the attitude one expresses, on the basis of that expression), a generalization of the notion of an assertion. I then argue that analogues of some classic norms of assertion, a class that I called *fittingness-based* norms, don't apply straightforwardly to some attitudes, like amusement and excitement. I then present an alternative, rightness-based theory that can capture what motivated the fittingness-based theories for assertion, while also not delivering the wrong results for these other attitudes. It is essentially an interest-based, contrastivist approach to these issues. I also try to account for norms for the use of what linguists sometimes call expressives.

CHAPTER 2

Policy Externalism

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 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]

2.1 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.).

[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)] calls *Russell's Principle*, namely that "a subject cannot make a judgment about something unless he knows which object his judgment is about" (89). (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.

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 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 $\mathbf{P}$ is *credence-involving* iff there is some rule in $\mathbf{P}$ that specifies that the agent must have some kind of credence or other concerning whether the circumstances of the rule obtain. Otherwise, $\mathbf{P}$ 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,

[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.

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 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 $X_1, ..., X_n$ across $w_1, ..., w_n$, 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

[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.

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 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 $\mathbf{P}$ for A-type attitudes is unreasonable to have if there are clearly better A-type policies than $\mathbf{P}$ 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

[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.

[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)].

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.

2.2 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 some problems that might seem to arise for that 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 $F_1, ..., F_n$.

b. Hate X just in case X kicked your dog.

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

(4) a. I hate whoever has properties $F_1, ..., F_n$.

b. I hate whoever kicked my dog.

Here's one from the wild:

(5) I HATE whoever invented TV commercials.[11]

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 $F_1, ..., F_n$, (ii) X bears A to every Z that

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

has properties $F_1, ..., F_n$ 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 $F_1, ..., F_n$, 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 $F_1, ..., F_n$, 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 resemble relational attitudes, though, in that they are *specific*: if Karin is the one who kicked my dog, then in virtue of subscribing to (4b), 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.

[12] See [Quine(1956)].

[13] (a) is something like '$\textsc{bel}(\text{Ralph}, (\exists x)(\textsc{spy}(x))$', and (b) is something like '$(\exists x)(\textsc{bel}(\text{Ralph}, \textsc{spy}(x)))$'.

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 $x_1, ..., x_n$ are such that x_1 has $F_1, ..., F_m$, ..., x_n has $G_1, ..., G_k$. X can conform to this policy by employing mental names whose references are fixed as in (8) and (9), thereby *unfailingly* bearing A to $x_1, ..., x_n$. 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)], [Kaplan(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

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

> 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 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]

[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)], page 263.

[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.

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.

(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 $\ulcorner X$ believes that N is $F\urcorner$ has the following form:

(16) $\textsc{bel}(X, \langle F, N\rangle, m)$,

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) $\textsc{bel}(X, \langle F, N\rangle, m)$,

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,

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

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 from my beliefs about the matter.[23] It is then no mark against our having a particular resentment against X that 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 reason-

[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.

[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)], page 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.

ableness 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.

What about P3? Here it is again:

P3. A policy $\mathbf{P}$ for A-type attitudes is unreasonable to have if there are clearly better A-type policies than $\mathbf{P}$ 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.

[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.

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.

2.3 The Linguistic Argument

2.3.1 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.

2.3.2 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

[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)]).

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'):

(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:

[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)], page 29 for this idea applied to the *Miners' Puzzle* (see below).

(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:

(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 $\ulcorner(\phi \wedge \psi) \Rightarrow (\chi \wedge \neg\psi)\urcorner$, 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.

[32] So-called because [van Fraassen(1980)] attributes them to Richmond Thomason.
[33] See, e.g., [Stalnaker(1968)] and [Lewis(1973)].
[34] See [Davison(1979)] and [Iatridou(1994)].

2.3.3 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.

(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 *modus ponens*.[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:

[35] See also [McGee(1985)].
[36] The case originates with [Regan(1980)], and it received prominent discussion in [Parfit(1984)].
[37] See [Kolodny and MacFarlane(2010)] for the simple derivation.
[38] See [Kolodny and MacFarlane(2010)] and [MacFarlane(2014)], chapter 11 for further details.
[39] See [MacFarlane(2014)], page 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.

(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. If they're in B, then 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, (31a–b) entails:

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

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:

[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, see [Moss(2013)], [Swanson(2016)], and [Yalcin(2012)].

(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) $[[\ulcorner$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) $[[\ulcorner S$ 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 (30a, b) are true, but the negation of (c) doesn't follow.[45]

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 the miners are in A, then Block wants to block A;

[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'.

b. and if the miners are 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 (37a, b) that sound all right in the stipulated circumstances. These readings call to mind examples like [Williams(1981a)]'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 *subject'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.

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

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.

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 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.

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

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

Normality. $v_X(W) = 0$.

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

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 $v_X(\{w\}|p) > v_X(\{w'\}|p)$ 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)Cariani, Kaufmann, and Kaufmann].

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 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

[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 $\phi \wedge \chi, \psi\urcorner$.

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 (37a–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 (37) true, it would not extend to the other attitudes in conditionals like (19)–(21); we've also seen reason to think that (37) 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.

2.4 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

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

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]

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:

[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.

(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 conditional 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.

CHAPTER 3

Wanting What One Should Want

Introduction

We spend so much effort day-to-day on coming to believe the things we should: on gathering evidence, on determining the right standards to apply, and on looking for overlooked possibilities. Imagine a friend tells you that you can lay down your burdens, because it's actually quite easy to figure out what to believe: you should believe what's true, and disbelieve what's false.[1] In a way, they said something true; I should believe what's true and disbelieve what's false. But that doesn't mean it's easy to figure out what to believe, since it's not easy to figure out what's true and what's false. That task is exactly the task that we spend all that effort on in the first place. Your friend has gotten you almost no further to the end of it.

We also seem to spend a great deal of effort day-to-day on coming to *desire* the things we should. Should I want this sort of life or that sort? Should I want to go to this school or that one? In this paper, I'm going to play a role not too different from the friend who tells you the whole thing is easy: you should want what you should want.[2] This will at first blush look even less helpful than the friend you imagined telling you what to believe. At least they took a stand on the issue, since though the claim that you should believe truths and disbelieve falsehoods might be trivial, it is not quite as completely empty as the claim that you should want what you should want. My aim is not to show that it's *true* that you should want what you should want, since I doubt anyone would disagree that you should. Instead I'll try to show how the advice can actually be conformed to in a way that makes it more helpful than the advice about belief was. As a corollary, I will argue for a limited kind of moral rationalism, according to which some desires are rationally required because they are morally superior desires to have. That argument will come later. I first need to say

[1] Exceptions might have to be made for the Liar, borderline cases of vague predications, etc.

[2] I will use 'want' and 'desire' interchangeably, adjusting syntax where necessary.

more about why I believe we can actually heed the advice.

My general purpose, though, is to try to illustrate an approach to rational desire, and thereby preference,[3] that differs markedly from all its competitors. That approach might be right and it might not be; my purpose will be primarily to illustrate how it works, and how it might overcome apparently strong objections. To that end, section 1 makes an initial case that we can conform to the advice, section 2 argues that the advice is substantive because it generates highly non-trivial moralized rational requirements, and sections 3, 4, and 5 formulate and address the strongest objections I could think of to the general approach on offer here.[4]

3.1 General Desires and General Beliefs

I'll first explain what I take to be at the heart of the difference between belief and desire that I just gestured at.

I need the notion of a *general* attitude.[5] Examples of general attitudes include hating whoever abuses dogs, admiring anyone with perfect pitch, and regretting whatever things you've done that badly hurt people. These attitudes can be a bit hard to place, and I won't be offering a theory of them.[6] I instead want to assemble some important observations.

Suppose my evidence leaves it open which element p, q, or r of a partition is true. Suppose I also know that one of p, q, or r would make me happiest were it true. In that sort of case, I can't simply believe whatever's true, even though I might take a leap and come to believe the proposition that is *in fact* true. But I can want whatever would make me happiest were it true to be true. There does not have to be a "leap" in that sort of case, either. Here's a different way to put things, now in the formal mode. In such a circumstance, I cannot appropriately say:

(1) I believe whichever of p, q, or r is true. So if p is true, then I believe p, but if q is, I believe q, and if r is, I believe r.

But I can appropriately say:

(2) I want whichever of p, q, or r to be true that would make me happiest were it true. So if I'd be happiest if p were true, then I want p to be true, but if I'd be happiest if q

[3] Assuming that S prefers x to y just when S desires x more than they desire y.

[4] One objection I don't address here is the one I address at the end of chapter 4 ("Deliberation, General and Particular"), that I am suggesting desires agents would largely be alienated from. I think what I say there gives the beginnings of an answer, but there's a lot to say about it.

[5] For some discussion as relates to the philosophy of language, see [Blumberg and Holguín(forthcoming)].

[6] See chapter 1 ([Drucker(forthcoming)]), section 2 for the beginnings of one.

were true, then I want q to be true, and if I'd be happiest if r were true, then I want r to be true.

For the time being, I want to be clear on what I am *not* saying. I am *not* saying that the conditionals in (1) and (2) need to be read as ordinary conditionals, the sort of thing you can use *modus ponens* on unproblematically. I think that, but I have not established that. What matters is just that we can desire whatever would make us happiest, say, even if we don't know what would make us happiest, but we cannot believe whatever's true, at least not without a "leap".

That observation has been put very abstractly, and in part imprecisely. So allow me to clarify what I mean by a "leap". The first sentence in (1) can sound appropriate, and so true, when accompanied by the right things. So, for example:

(3) I believe whichever of p, q, or r is true, because I believe p, and p is true.

The speaker clarifies: they do not have an "ungrounded" general belief, but a general belief *grounded in* a particular belief, namely the belief that p. They believe whatever is true, they think, because they believe p, which is true (they think). If they could not point to one of p, q, or r as the thing they believed, then they would not be able to say (1) or (3) appropriately, and thus it seems, not truly. That is not true of other sorts of general attitude, and in particular it is not true of general desires. To see that it is not true of other sorts of general attitude, imagine I know one of Alice, Bert, and Carlos abuses dogs, and nothing else relevant about them. I can hate whichever of them actually does abuse dogs. Because of that, I can say:

(4) I hate whoever the dog abuser is.

I do not need to add the following:

(5) I hate whoever the dog abuser is, because I hate Bert, who is the dog abuser.

I could add that, but it is not necessary to do so. Similarly, I could say:

(6) I want whichever of p, q, or r to be true that would make me happiest were it true, because I want q to be true, because q would make me happiest were it true.

(6) is perfectly fine, but I don't need to add it to what I say in (2). There are "ungrounded" general hatreds and desires, whereas there do not seem to be "ungrounded" general beliefs. To put it another way, if I believe whichever of p, q, or r is true, then in normal cases I need

to be able to *point to* which of p, q, or r I think is true. I say "normal cases" because in cases where my beliefs are not luminous, i.e., where I believe that ϕ but do not know that I believe that ϕ,[7] I might still believe whichever proposition is true, but not be able to point to the one I really believe is true. Such cases won't be relevant to the main thread of this paper, though, so I will ignore them.

Here is the observation, then:

> **Observation 1**. We can have general desires for whatever would make us happiest without being able to indicate which things would. We cannot have general beliefs about which of a partition of propositions is true without being able to indicate which proposition is true (beyond just by using 'the true member of the partition' and similar descriptions).[8]

Perhaps there are descriptions with which we cannot form general desires. Perhaps we cannot have desires for what is already true, for example, or that we believe to be already true.[9] So far, the observation is quite limited. It is still important, though. I started with the idea that "believe what's true and disbelieve what's false!" is unhelpful advice, because not directly actionable, whereas "desire what you should desire!" might be better. **Observation 1** explains the first part of that: belief in whatever's true cannot be ungrounded. We've also seen reason to be optimistic about desire, since I can have the ungrounded desire for whatever would make me happiest. But we are still a long way from our goal.

Say that a description D *supports* a person S's bearing an attitude Φ to whatever satisfies it when S can have an ungrounded Φ-type general attitude canonically ascribed by means of D. 'True member of a partition of propositions' does not support belief, but 'something that would make S happy were it true' does. I did not pick those descriptions randomly; that something's being true would make a person happier *gives them reason* to desire that thing. It is a fitting reason to desire a thing, just as something's being true is a fitting reason to believe it. Can we make that connection even tighter for desire, between descriptions that support desire and that provide strong, even decisive reasons for the desire?

That something's being true would make you happy is a relatively "thick" reason for you to want that thing to be true. Because it is thick, it is less likely to be decisive. Perhaps a lot of things trump happiness, especially one's own happiness. That is indeed frequently

[7] See, e.g., [Williamson(2000)].

[8] I chose to formulate it this way rather than a more general way I could have. The points about 'true member of the partition' apply to whatever other descriptions you like—one still needs to indicate which propositions satisfy the description. But other descriptions are not relevant to what I say.

[9] Socrates assumes this in the *Symposium*; see also [Lauria(2014)].

the case: I shouldn't want half the world I'm isolated from to suffer, just so I can be happier than if they didn't. Or, if you think I'm begging some questions in saying I *shouldn't* want that, then at least I am not *required* to want it. Imagine there are two candidates in an upcoming election, A and B, but I've been too lazy to figure out what each supports, *a fortiori* what would be the best policies to support. I can still want whoever would do the better job to win. Indeed, *ceteris paribus*, I *ought* to. I just won't know which of them is which. In that case, 'whichever candidate would do the better job' supports my desire. That is a much thinner description than before, since 'better' is compatible with all sorts of things being good and bad. Yet it is still relatively thin. First, there are cases in which it is better to hope for the candidate who would do the worse job, if for example accelerationatism is called for in those circumstances.[10] Second, you might have made and forgotten a promise to support B; that promise might overrule a not-too-drastic difference in governing ability. Nevertheless, this should make us even more optimistic, since 'whoever would do the better job' is so thin.

How thin can descriptions be and still support desire? Very thin, it turns out. In figuring out how to come a fair agreement with an ex about how to divide things up post-breakup, I can want whatever my fair share is, for example. Or when faced with a morally difficult choice, I can want to do whatever's right. (It's another matter whether I *should* want to, in this general kind of a way.[11]) Another way I might put that is: I want to do what I ought to do, or what I should do. 'To do what I should do' is an incredibly thin description, and yet it still seems to support my desire. There seems to be no difficulty of principle that these thinner descriptions introduce. Thus:

Observation 2. Very thin normative descriptions like 'to do the right thing' and 'what's best' support desire.

We have now worked or way up to the central issues of this section. Does anything change when our descriptions are of the form 'whatever it's F to desire'? In discussing these kinds of description at all, we have, at least, departed from ordinary language quite a bit. Start again with other attitudes; I'll turn to desire in a moment. Here's the kind of self-ascription I have in mind:

(7) I hate whoever kicked my dog, so long as it's not wrong to hate them. (That is, I hate whoever kicked my dog that it's not wrong to hate.)

[10] According to Wikipedia: Accelerationism, it's "the idea that either the prevailing system of capitalism, or certain technosocial processes that have historically characterised it, should be expanded, repurposed, or accelerated in order to generate radical social change".

[11] It might make me a "moral fetishist"; see [Smith(1994)] and [Arpaly(2002)] for interesting discussion.

My ear hears nothing wrong with (7), though perhaps it has been corrupted by theory. You might wonder why someone would say or think it. Kicking my dog is egregious, but there are an open-ended number of things that would make it wrong for me to hate such a person. They might have just lost a parent and kicked him in extreme despair, for example. In saying or thinking (7), I express my desire not to be a jerk. On that note, take the following:

(8) I regret whatever I've done I would have to be a jerk not to regret.

Of course, not being a jerk comes to more than making these utterances true. The regret shouldn't be thought real if, even on being convinced that one would have to be a jerk not to regret having done particular action, one is still proud of having done the thing. But one *can* be sincere in uttering or thinking (7) and (8). If there is some large difference between (8) and (9), I cannot think of it:

(9) I regret whatever I've done that was horrible or cruel.

With hatred and regret, at least, we can have general attitudes that involve descriptions that themselves involve hatred and regret.

Finally, turn to examples like this:

(10) I want to be whatever would make me happiest that wouldn't be too selfish to want.

Similar to (7) and (8), someone who says (10) has a genuine desire for their own happiness, tempered by not wanting to desire too selfishly. This is a completely understandable desire. It might even be virtuous. In other words, descriptions involving desire *can* support desire. That is what the rest of our examples should have led us to expect, too. It is hard to see what would have prevented examples like (10) from making sense or possibly being true.

Thus, the final observation I will make in this section:

Observation 3. Descriptions of the form 'whatever it's F to desire' can support general desires.

I've argued for the following claims:

- We have principled reasons for finding the advice "believe what's true and disbelieve what's false!" unhelpful, even if other descriptions support other general attitudes.
- General desires are supported by very thin normative descriptions.

- These descriptions might include mention of the very attitude supported by the description, including desire.

Together, that means there is no in-principle reason to think that the description 'whatever I should want' cannot support desire. That is, we have not found any in-principle reason to think that we cannot comply with the apparently unhelpful advice from the introduction. That does not mean that we *can* heed it, just that we have more reason to think we might.

Before I continue on, I want to return to an issue from before. Recall (2):

(2) I want whichever of p, q, or r to be true that would make me happiest were it true. So if I'd be happiest if p were true, then I want p to be true, but if I'd be happiest if q were true, then I want q to be true, and if I'd be happiest if r were true, then I want r to be true.

I said before that I did not want to take a stand then on whether or not we can use *modus ponens* on these conditionals. If we could, then supposing p would make me happiest, it would turn out that I would desire that p be true. For example, if I would be happiest as a lawyer, then the following would be true of me (though I wouldn't know it):

(11) I want to be a lawyer.

I also said I believed we *could* detach. My reason is that we just do detach, in ordinary life. It's seamless when we do it about our own desires, and so can't show much. But take cases like [Williams(1981a)]'s, in which a person thinks a glass of gasoline is a glass of gin. If we know it, we can say 'you really don't want to drink that!' We can say it because we know the person has a general desire not to drink poisons. These cases proliferate; when we take our interlocutors' interests and their ignorance to be obvious, we often feel comfortable attributing those desires to them. That is part of why I think we should say that general desires, with the facts, entail specific desires. More importantly, but perhaps less dialectically usefully, the linguistic forms themselves just clearly do support those entailments, at least in other contexts. But I won't press that point.

Given this, it would be strange for me to have any *other* desire about what career to have. Suppose, for example, that though I would be happiest being a lawyer, I *think* I would be happiest being a doctor. On that basis, I come to desire being a doctor, while still desiring being a lawyer. In worlds like these and for people like me, those are not jointly satisfiable desires. (We can get worse incoherence easily, but the point is clear.) There is simply no reason to add these desires.

This contravenes certain theories of rational "instrumental" desire. The thought is supposed to be that I should prefer p to q if p's truth brings with it more of what I intrinsically desire than does q's truth.[12] Given that I desire whatever career would make me happiest, it seems confused to suggest that I *also* ought to want to be a doctor. This confusion is compounded if the theory tells me that I ought to *prefer* being a doctor to being a lawyer. We have no need of a theory of rational preference like that, tied as they are to what might be very inaccurate subjective probabilities. Such theories of instrumental desire are better understood, I think, as theories of rational intention or action. While I shouldn't prefer to *be* a doctor, I should (subjectively) take actions that would lead to my being a doctor. I will return to these issues in a little while.

The upshot of all this discussion is that there seem not to be reasons for me to desire anything *besides* whatever I should desire, all things considered. And if my observations are correct, then there are no in-principle reasons yet for thinking that I can't desire in that way. My argumentative burden is not, however, to show that we can desire what we should desire. That much might even be trivial, following as it does from a version of 'ought' implies 'can'.[13] I want to show how, strange as it might be, the advice to desire whatever one should desire, all things considered, is correct and non-trivial. Even if it is easy to heed that advice, it is also easy not to. In the next section, I will nail down exactly what the description I have in mind is, and then I will illustrate why the advice it gives is non-trivial.

Finally, you might wonder whether any of this is specific to desire. I have tried to be explicit about it: I do not believe that it is. I think that whatever makes various descriptions support desire will make them support hatred, regret, and the rest. In part, my project is an experiment in applying these ideas to a specific attitude. But in the next sections, more about desire in particular will be relevant.

3.2 Why the Advice to Want What One Should Want Is Nontrivial

The most important part of section 1 as far as my discussion going forward is concerned is that descriptions with normative terms and 'desire' can support desire. I will assume that that's right. The description I will focus on is this:

$\mathcal{D}$. 'whatever I objectively should want, all things considered'

[12] See, e.g., [Jeffrey(1965)] and [Joyce(1999)].

[13] I say *might* be trivial; the principle is contested, so I don't assume it. See, e.g., [Graham(2011)].

Before moving on, I will clarify two parts of $\mathcal{D}$, 'objectively' and 'all things considered'.

Though I use 'objectively', the description is also meant to generate a *subjective* 'should'. In other words, since people know that what answers to $\mathcal{D}$ are, trivially, the desires they objectively (roughly, independent of their beliefs) should have, they thereby also subjectively should have those desires.[14] When one knows how to conform to what is objectively required, one then subjectively ought to conform—they are blameworthy or otherwise criticizable if they don't.[15] Ultimately, I needed to put things in terms of the objective 'should' so as not to beg any questions against rival theories of rational desire; it could be that if $\mathcal{D}$ had 'subjectively' in terms of 'objectively', then the approach here would be extensionally equivalent to what I take to be its competitors.

As for 'all things considered', I mean to rule out "partial" desires. We sometimes use 'want' when we see something good about a thing, even if on the whole we don't pursue it because of some other bad features associated with getting it. I want to go to the gym, for example, but I don't want the hassle of transit. So, on the whole, I might not want to go to the gym, all things considered (in particular, the unpleasantness of the trip). The restriction to all-things-considered desires is not crucial, but it does simplify the discussion by a lot.[16]

Here is the first issue. Some definite descriptions, including plural ones, do not refer. As I write this, 'the people who have read a complete draft of this paper' is non-denoting. So, why think that $\mathcal{D}$ denotes? After all, according to linguists, -wh-ever phrases ('whatever I should want', 'whoever would do the best job', and so on) are definite descriptions, at least semantically.[17] They thus have the truth-conditions of ordinary plural definite descriptions.

When, then, do plural definite descriptions denote? The answer is controversial. According to [Brogaard(2007)], a sentence of the form '$G[\iota X: Fx]$' is true iff $(\exists X)(FX \wedge (\forall Y)(FY \supset Y \subseteq X) \wedge GX)$, i.e., when there are some things that are F and if some (possibly different) things are F, then those second things are all some of the first things—and those first things are G.[18] So take the following:

(12) I want whatever I objectively should want, all things considered.

On the present view, that sentence is true when there are some things I should[19] want, and possibly some other things I should want, and all those second things are some of the first

[14] For the distinction between subjective and objective 'ought's (and thereby 'should's, which I use here interchangeably), see, among many others, [Gibbard(2005)], [Carr(2015)], and [Wedgwood(2014)].

[15] Would this work the other way around? If you put 'subjectively' in $\mathcal{D}$ instead of 'objectively', does $\mathcal{D}$ then generate an objective 'should'? The issues are tricky, turning on just what beliefs a person should have. (In part, should they have gone through the reasoning of this paper?).

[16] For further discussion, see, e.g., [Davis(1984)], [Levinson(2003)], and [Villalta(2008)].

[17] See, e.g., [Jacobson(1995)].

[18] Essentially, Brogaard reframes [Sharvy(1980)]'s analysis in terms of plural variables and quantification.

[19] I will sometimes drop some of the other verbiage.

things—and I want those first things. There needs to be a set of all the things I should want. That means need to be things that I objectively, all things considered should want.

Are there such things? I think so. Imagine a person who has experienced a great deal of hardship in their life, but who throughout has always maintained a cheerful and kind disposition and has worked hard to make the best of rough situations. Call this person Job, and assume further that no changes to Job's well-being will have physical effects on any relevant individual (including our potential desirer S, but also anyone else you like—his being made better off, say, won't have knock-on effects that lead to mass misery). Suppose, finally, that S's desires *are* somehow causally efficacious for Job, in that the more S desires that Job not be bad off, the less bad off Job will be. (Imagine, if it helps, that S has a Job-focused wish-granting fairy granting their desires.) I claim that for any S in Job's world, S objectively should desire that Job be better off, all things considered. To *not* desire that nontrivially would be horrifically immoral, and not in the least offset by any prudential gain. After all, Job will never physically affect S, no matter his well-being; he is cheerful, kind, and hard-working, and so by any plausible theory of desert *deserves* to be better off than he is; and desires for his increased well-being will be *satisfied*, and so not frustrated or disappointed.

As I said, for people who live in Job's world, it seems as clear as anything could be that everyone should want Job to be better off than that. I am, in other words, confident in claiming that everyone in that world objectively should have that desire, all things considered. As long as the moral has *some* normative force, and as long as the other things that have normative force—prudence, aesthetics, etc.—are not very different from what they seem to be, then my claim about Job simply follows. This raises a puzzle, though. I said before that $\mathcal{D}$ generates *subjective* 'should's, the kind of thing a person would be criticizable for not complying with. Anyone in Job's world, I just claimed, subjectively should desire that he be better off. Indeed, I think it is *irrational* not to do what one knows one objectively ought to do, all things considered.[20] That would mean that any individual in Job's world is rationally required to desire that Job be better off. Since that is true ultimately because that desire is so morally obligatory, the view on offer supports a kind of moral rationalism: *pace* Hume, there are desires we are rationally required to have because they are so morally correct. That result is not at all trivial. In other words, the advice's being good advice would have a surprising normative upshot, that certain morally obligatory desires are rationally required.

[20] There are some tricky issues in the vicinity; what if we objectively should be irrational, and somehow know that? Are we then rationally required to be irrational? See [Schelling(1960)] and [Parfit(1984)], chapter 1 for discussion. Luckily such cases are very much not relevant to this discussion, so I will ignore them going forward.

Other theories that propose alleged rational requirements on desire do *not* get this result. Famously, decision theory of any variety does not. If someone does not start with moral intrinsic desires, then no standard decision theory will advise forming moral preferences because they are moral.[21] This has sometimes been taken to be an objection to such theories, at least as theories of rationality.[22] I confess I have never found such criticisms very persuasive. Theories of rational preference seemed acceptably garbage in, garbage out: that we shouldn't expect such a theory to give substantive rational requirements besides those of a kind of coherence at the level of preference.[23] An amoralist or immoralist need not have any moral desires, you might have thought. But now there seems to be something right about the objection. A thoroughly de-moralized theory of rational desire, that is a theory that does not recommend *any* desires because they are especially moral, is incorrect, at least if the approach on offer is correct.

Here's a different way to understand what I've been arguing in this section. In explaining externalism's differences with internalism, [Stalnaker(2008)] claims that much of it comes down to what the acceptable starting points are. The internalist takes internal facts as unproblematic, and tries to reason to how we can know and what we can know about the external world; the externalist does the reverse. I am appealing to a similar kind of externalism here. We can take our sense that the moral matters for granted, and that it helps determine what we objectively should desire, all things considered. I fully recognize that would not convince the amoralist or immoralist. They will simply deny that the moral plays that kind of role in determining what we objectively should desire, all things considered. Luckily I'm not trying to convince the moral skeptic; I am merely showing how those of us who are not moral skeptics can know that even the moral skeptics among us are rationally required to have certain desires, simply because they are objectively moral.

Before moving on, I want to emphasize something: Job's case was extreme, because I wanted to give a case in which there is no room for doubt. I suspect we could relax many of those assumptions, and come up with desires that are no less rationally required for all that. I won't explore that suspicion any further in this paper, though.

If this section's argument is correct, I have shown that the advice in the introduction was, surprisingly, non-trivial. It gives different advice from decision-theoretic theories of rational preference, in particular to have certain desires simply because they are the morally correct ones to have. In the next three sections, I'll address three objections to the general approach. The first concerns its normative adequacy, asking whether it leads to unnecessary

[21] They might for other reasons, like that morality happens to pay, or even necessarily pays.
[22] See, e.g., [Quinn(1992)].
[23] I'll come back to coherence of preference in a little while.

incoherence. The second asks how, given how I address the first objection, the view can be substantively different from other approaches in ways that matter. The third asks how we really can possibly have the desires I say we can have.

3.3 Ignorance, Incompleteness, and Incoherence

It was important to the argument of section 1, and formed the basis of **Observation 1**, that we are often ignorant of what we bear our general attitudes to, specifically. I may hate whoever kicked my dog while having no idea that it's Bert that I hate. This ignorance can be a strength: as I tried to show in the last two sections, we can have especially good desires that way. But it can also be a weakness. If we don't know what our general desires are desires for, we can end up having incoherent desires, roughly through no fault of our own. So, imagine someone—call her Maggie—who takes my advice in the introduction and comes to have those desires she objectively should have, all things considered. She thereby comes to have certain desires that she doesn't know that she has, let's say that she become a journalist. She also realizes that $\mathcal{D}$ is likely to give her a very incomplete desire set: she thinks there are certain states of affairs and complements, neither of which she *should* desire obtain—she is free to pick, as it were. There are a number of ways this could happen. It could be, for example, that pluralism is true, that is there are multiple different normative theories that assign states of affairs overall value, and the world itself doesn't choose one among them as *the* right one. Or it might be that there are cases in which, for any given state of affairs of a given type, there is a better one, such that there is no single state of affairs she should want to obtain.[24] The point is, the supererogatory is likely to be rampant in this area. So to have anything like a complete *enough* desire set, we would need to supplement the general desire for what one should desire with other desires, perhaps general, perhaps not. The trouble is that, because Maggie will be massively ignorant of what she desires by means of $\mathcal{D}$, in supplementing those desires, it will be merely adventitious if she avoids incoherence.

By 'incoherence', I mean what decision theorists have meant. Desires have objects and strengths; together, they induce a preference ranking: S prefers x to y iff S desires x more strongly than they desire y. A preference ranking is coherent iff it is corresponds to or can be extended into corresponding to a *utility function*, a function from the objects of preference to real numbers unique up to positive affine transformation. Here's the kind of case that can lead to incoherence:

[24] For this thought as applied to the problem of evil, see [Plantinga(1976)].

Journalist or Politician?. Maggie is clearly cut out to be a journalist, and it turns out she would definitely do the most good were she to become one. Let's suppose, then, that she objectively should want to be a journalist, all things considered. On the other hand, Maggie was antecedently pretty sure that one's career choice was exactly the sort of decision she'd have a wide range of latitude on. She forms the desire to have whatever career would make her family proudest, and desire to not have whatever careers would not make her parents proud. It turns out that being a Congresswoman is what would make her family proudest.

If Maggie forms her desires in line with those descriptions, then she will desire both p and $\neg p$, i.e., to become a journalist and not to become a journalist (but rather a Congresswoman instead). If to desire that p (all things considered), though, is *at least* to prefer p to $\neg p$, or to prefer the world that would be were p true to the world that would be the case were it not, then Maggie would have asymmetric preferences: she would prefer p to $\neg p$, and $\neg p$ to p. Those preferences cannot be part of a set of preferences that corresponds to a utility function. Thus, Maggie would be incoherent. This situation leads to the following objection:

Objection 1. Forming desires as I've suggested makes avoiding conative incoherence a matter of luck. It is irrational to form desires in ways that might easily lead to incoherence, if there are theories that can avoid that result. There are: the different varieties of decision theory.

I think we should not be worried about incoherence like this. Our best theories of choice as opposed to preference ought to lead us to make coherent choices, but since these sorts of incoherence arise exactly because and when our agents are ignorant of exactly what it is they prefer, they will not lead directly to incoherent choice behavior. The rest of the section will elaborate that basic response.

Objection 1 belongs to a general class of objections to theories that posit unlucky incoherence. The most famous comes from [Kripke(1979)]:[25]

Puzzling Pierre. Pierre was raised in a small French village as a monolingual Francophone. Based on movies, advertisements, etc., he came to accept the following sentence:

(13) *Londres est jolie.*

[25] I am, of course, simplifying Kripke's very rich example and discussion. See that paper for details!

Eventually he moves to England, specifically an ugly part of London unrepresented in the materials of his youth. He then comes to assent to the following sentence, not knowing that '*Londres*' and 'London' co-refer (he learns English by immersion):

(14) London is not pretty.

We seem to have strong reason to say that Pierre believes a *proposition* and its negation, that London is pretty. But Pierre seems guilty of no logical error or other kind of irrationality. He seems to be incoherent merely because unlucky. (Assume no fancy ways of distinguishing the two beliefs so that they don't *really* conflict succeed.[26]) Now suppose an epistemologist were to come along and say that theories of rational belief formation ought not to allow beliefs to be formed in ways that might as a matter of misfortune lead to doxastic incoherence, i.e., where one's beliefs entail p and $\neg p$ for some p. She says that Pierre *did* make a mistake, but not a logical one. His mistake was using the wrong kind of sentences to form his beliefs. He would have done better to have formed his beliefs with the following sentences:

(15) The city I associate with such-and-such movies, advertisements, etc. of my youth that I take to be called '*Londres*' is pretty.[27]

(16) The city in which I now live that I take to be called 'London' is not pretty.

Given what *we* know, we know that Pierre is still guaranteed to have a false belief. In other words, his actual accuracy will be unaffected by the change. That's fine, though, because he will not be doxastically incoherent. It's logically possible for (15) and (16) to be true together.[28] So, should epistemologists recommend against forming beliefs with proper names? (If so, should they also recommend against forming beliefs with sentences with terms they might have incomplete mastery of?)

In the general case, the answer seems to be: clearly not. Perhaps refraining from forming beliefs in that way would save us from the kind of unlucky incoherence to which Pierre seems subject. But what great service is that? He's still just as wrong in *fact* as he always was. The Pierre who believes the propositions expressed by (15) and (16) is not doing better in any meaningful sense than the Pierre who believes the propositions expressed by (13) and (14). The incoherence the more punctilious Pierre avoids does not seem *worth* avoiding. He is just as wrong, nor will his actions be any better. Either way he'll try to

[26] I think [Burge(1979)]-style 'arthritis' beliefs can also generate such cases, but I won't pursue that.

[27] Or whatever the French translation of this is.

[28] As I read him, [Lewis(1981)] actually has the beliefs expressed by (15) and (16) rather than the ones expressed by (13) and (14). I think that's wrong, but arguing against it would take me too far afield.

leave the city he then lived in that he finds pretty, and to visit the one he remembers from his childhood. We must be wary of coherence fetishism.[29] In fact, it seems like the Pierre who believes with (13) and (14) is better off than his counterpart, insofar as tags are easier to work with than complicated descriptions. That advantage is probably minor in his case, but minor advantages can be large ones over a lifetime. *Knowable* incoherence is bad, most likely; but unknowable incoherence doesn't seem to be worse than equal actual inaccuracy.

I think something similar is true of Maggie. Incoherence in preference can lead to imprudent choices, for example when intransitive preferences leads one open to money pumps. But the same ignorance that leads Maggie to incoherence will also lead her to not make bad decisions. She won't, for example, constantly switch between pursuing journalism and politics for small fees. That's because she won't *know* she both wants to be a journalist and not to be a journalist but rather a Congresswoman. Given how I've stipulated her views, she'll try to pursue the career that would make her parents proudest, since she is pretty sure (wrongly, it turns out, but let's say not unreasonably) that within a certain range of careers, it is all right to want to pursue any one of them. Even though she would then go against another desire she has, namely to be a lawyer, she doesn't know that, and so won't be led to the kind of incoherent choices that make one susceptible to money pumps and all the other devices to exploit incoherence.

Let's do things a bit more carefully. For simplicity's sake, assume there are two dimensions along which we can evaluate desires: how good, morally correct, etc. the actions they lead to are; and how fitting they are. In discussing Pierre's beliefs, we applied two similar dimensions of evaluation: first, how good, etc., the actions his beliefs lead to are; and second, how fitting, i.e., accurate, evidentially warranted, reliable, etc., his beliefs are. Consider the first dimension first. I said before that I thought standard decision theory was susceptible to counterexample as a theory of rational desire, but that doesn't mean I don't think some version isn't right about actions themselves, or intentions. I would like to say that we should simply graft such a theory onto the theory of rational desire I've been discussing. But that's tricky. Very probably $\mathcal{D}$ does not just stand for instrumental desires, but also some intrinsic ones too. That means that very likely on the approach I'm discussing we'd be very ignorant of much that we intrinsically desire. If we apply standard decision theory to intention, we would have a theory that says something like: intend to perform the action that has the largest expected intrinsic utility, where the expectation is determined by how likely you think the action makes each outcome multiplied by the intrinsic utility each outcome generates for you, summing over all the possible outcomes. It's hard to come close to knowing what maximizes that expectation when you don't know what you intrin-

[29] [Kolodny(2007)] comes to a similar conclusion but with very different motivations.

sically desire, at least not in a usable form. We could, then, apply expected *expected* utility theory: the expectation would then be calculated with all the intrinsic utility functions that might be yours, weighted by how likely you think they *are* yours.[30] But then again, why think that even on the standard approach we *aren't* pretty massively ignorant of what we intrinsically want? Recent work on transformative experiences argues that we are, at least for many very important decisions.[31] But we don't even need to go to that work to find it plausible that we are very ignorant of our intrinsic utilities; we are often surprised to find ourselves finding a piece of news welcome or not, we don't know whether we love or hate a person (and thus desire their welfare intrinsically), etc. So perhaps we don't need to say much about how to adapt decision theory for intentions to cohere well with the present approach.

The point is just that my approach can rationally require us to intend exactly the same things that standard varieties of decision theory can. If we want to evaluate that approach by the actions it recommends, then it need not differ from what those other theories give. If conforming to those other theories rules out problematically incoherent choice behavior, then my approach can rule that out, too. Just as the two Pierres I presented would perform identical actions (besides, perhaps, verbal assent to very specific sentences), agents who follow the approach here need not act any differently from their counterparts, either.

The other dimension you might evaluate Pierre's beliefs along was their fittingness—truth, accuracy, etc. As I said, the two Pierres will not differ from one another there, either. Here, though, agents who follow my approach might very well differ from agents who follow a competitor. But I don't think those differences will tell against my approach. What fittingness for desire comes to is just a little obscure; it might be the object's goodness, it might be something else.[32] Either way, if Maggie conforms to $\mathcal{D}$, then she will have a bunch of fitting desires. It might be she also will have unfitting desires. But there is no in-principle reason to think the rest of her desires will be *less* fitting than rival approaches. Just as the Pierre who believes what he believes by means of (13) and (14) is no less accurate (or responsive to his evidence, etc.) than the Pierre who believes by (15) and (16), I can't see why incoherence as such would make Maggie's desires less fitting. The incoherence does likely entail that some of her desires are unfitting; but I see no reason to think that it adds to it.

In sum, **Objection 1**, the objection from incoherence, should not cause someone otherwise sympathetic to the present approach to worry. But something in how I responded to

[30] See, e.g., [Boutilier(2003)].
[31] See especially [Paul(2014)].
[32] See, e.g., [de Sousa(1974)].

that objection in this section might. I said that, regardless of what theory of rational desire and preference we use—the one under discussion here, or whatever else—we can *still* use decision theory to figure out what we should intend, choose, or do. You might wonder whether anything substantive hangs on the choice, then. In the next section I will argue that substantive issues do in fact turn on the choice.

3.4 How the Choice is Substantive

I said before that there was a substantive difference between the approach this paper presents and its decision-theoretic rivals (among others), in that, e.g., this paper's approach says that certain desires are rationally required because they are especially moral. But some of the things I just pointed to in the last section might seem to rob the dispute of ultimate significance. This section's objection, then, is very easy to state:

> **Objection 2**. If the approach under discussion is compatible with a decision-theoretic treatment of rational intention, choice, or action, then the choice between this approach or its rivals is not substantive.

The choice is substantive because, at least on a number of plausible theories, what desires we have has knock-on effects independent of what we do. Ultimately many have to do with welfare, but there are other things, too.

When Plato's Socrates proposes the approach that inspired this one in the *Gorgias*, that we all desire just whatever's good,[33] it was in the service of arguing against a conception of power that his interlocutor Polus had. If we think of power as doing what one wants, as opposed to what one *thinks* one wants, then someone who thinks they want something bad but doesn't actually will not be exercising their power in obtaining or realizing the bad thing. This combined with Socrates' view that everyone wants only whatever is good means that many of the tyrants who seemed to exercise power, then and now, didn't.[34]

Though we, or at least I, am not so interested in whether we should say that tyrants exercise power, there are connected issues that seem more important and striking. Suppose desire satisfaction is, if not the whole, at least a *part* of welfare. That is, suppose *ceteris paribus* that a life goes better to the extent that it includes fewer frustrated and more satisfied desires (or perhaps just intrinsic desires). Then which desires we have, or are rationally

[33] At roughly 468c and surrounding. I think there's a good case to be made that he means we're obliged (whether rationally, ethically, or some other way) to have those desires, but that doesn't matter to what I'm saying here.

[34] For a good discussion of this line of thought, see [Penner(1991)].

required to have, will make a difference to our welfare independent of whether we act any differently because of those desires. If the argument of section 2 is correct, then assuming that desire satisfaction does make a difference to our welfare, our welfare is in part moralized. That is, we are rationally required to be such that our welfare depends on whether certain morally relevant things happen, because they are morally relevant. This, too, was the sort of result that the ancients (Plato and Aristotle, particularly) wanted for welfare.

Another issue worth examination concerns autonomy: some have thought that one acts freely, or autonomously, etc., just when does *what one (really, truly) wants*.[35] Whether in φ-ing S acts freely or autonomously, then, will depend on what one desires, *even if* S would have φ-ed even if S had the desires recommended by rival approaches. This thought can even be parlayed into a defense of the Socratic idea that people do not do wrong willingly, but to do so would take a lot of work I cannot do here.[36]

This is a very partial list of the ways in which differences in what we desire can matter even if those differences do not result in differences in intention, choice, or action. There's a lot more to explore. What really matters, though, is that there *are* differences in desire that matter in this way, or at that least a good number of popular philosophical theories have it that there are. That's enough to answer **Objection 2**, even though a lot of work remains in exploring these topics and establishing the specific claims that some of these differences that matter rely on. In the next section, I will turn to the last major objection: can we *really* desire in line with $\mathcal{D}$? What would that even look like?

3.5 Having the Desires We Should Have

In the first section, I argued that there are no in-principle reasons to think that normative descriptions that mention desire itself can't support desire. That was the essence of **Observation 3**. That is, unfortunately, very different from giving positive reason for thinking that we can in fact have those desires. Yet everything depends on having positive reason to think that we can. The objection is thus simply put:

Objection 3. Why should we believe that we can have desires in line with $\mathcal{D}$?

It is simply put, but frustratingly difficult to answer, because it is not at all clear to me what the general form of an answer to it would look like.

[35] In addition to Hume, see, e.g., [Frankfurt(1971)].

[36] For a discussion of the relation between Socrates' claims that everyone desires the good and that no one does wrong willingly, see [Kamtekar(2006)].

Suppose someone asks you, *can* we want to go the store to get some milk in the next couple of minutes? I want to say that we manifestly do have desires like that. I can say that we definitely can because we just do. If we know anything by introspection about our desires, we know that. (Compare [Vendler(1972)], page 50: "As it is silly to ask somebody, 'How do you know that you are in pain?' it is equally foolish to ask, 'How do you know that you want to go to the movies?"') Well, here's one thing to say. S wants x just in case if S pursues x[37] and when S thinks they've gotten x (or x is realized, etc., depending on the ontological category to which x belongs), then *ceteris paribus* S is affectively satisfied. Those aren't infallible criteria, of course, but they are a start. To recall an example from section 1, if I want whichever candidate will do a better job to win, I'll work hard to form good beliefs about who would do a better job, and I will be happy, at least to some extent, if the candidate who wins is the one I am more confident will do a better job. That is compatible, of course, with wanting who I *believe* will do the better job to win, or who has the better expected performance. How might we confirm one over the other of these hypotheses?

We can certainly distinguish them introspectively. If some angel were to come to me and ask me whether I preferred (*i*) they make it so the people I believe would be the best people to do job J_1, J_2, etc., or (*ii*) they make it so the people who would in fact be the best people to do J_1, J_2, etc., I wouldn't hesitate in picking (*ii*), at least insofar as I trusted their judgment more than my own. Of course, as soon as the deal as struck, I'll come to *believe* that the people the angel picked are better than the people I thought best before, so the *ex post* satisfaction test will be tricky to make work. But the case seems clear even without checking for satisfaction.

Here's another way to argue for the same conclusion. [Evans(1982)] argues that we know what we believe by following or trying to follow the following rule:

BEL. If p, believe that you believe that p.

For example, if I'm trying to figure out whether I believe that dogs bark, I should ask: do dogs bark? If the answer I come to is 'yes', then I do indeed believe that dogs bark. Even trying to follow this rule works, since you apply it just when you think the antecedent is true. We need no special inner sense, nor need we look at our own behavior. [Byrne(2011)] has proposed the following rule for desire, meant to apply similarly:

DES. If p is desirable, believe that you desire that p.[38,39]

[37] Compare [Anscombe(2000)], that the "primitive sign of wanting is trying to get".

[38] Or *mutatis mutandis* for other potential objects of desire.

[39] See also [Moran(2001)]. [Ashwell(2013)] criticizes both Byrne and Moran, replacing DES with a version

For example, in trying to figure out whether I want to get some milk right around now, I should ask myself: is that desirable? If the answer I come to is 'yes', then I do indeed want to get some milk right around now.[40] I tend to agree with Byrne that DES is a good rule for figuring out what we want.

Now ask, do I think it's desirable that the candidate *I believe* will do a better job win, or that the candidate *who will* do the better job win? I've been misled and disappointed by candidates before, and so don't take myself to be especially reliable on the subject-matter. Because of that, I think the latter more desirable. And when I have to choose between those two options, I find the former, i.e., (*i*) not at all desirable as compared to (*ii*). What's attractive about (*i*) is, for me, screened off by what's attractive in (*ii*). So using DES I come to believe, and I think know, that I desire that the candidate who will in fact do the better job win.

Or, to change the case a bit, it might turn out that I didn't want the better candidate to win, because I have personal or affective loyalties to the other candidate that I wouldn't drop even on being fully convinced that they would do a significantly worse job. Avowing an attitude is not the same as having it. Indeed, *sincere* avowal is insufficient as well, since we can be misled about what attitudes we really have. Sometimes the only way to tell if what you desire really is that whoever the better candidate win, or that your friend win (who you think is the better candidate) is to be convinced that your friend is the worse one and see what you go for and how you feel. Even then, you might just *both* want your friend to win and the better candidate to win. These things can be incredibly hard to know once and for all, and in some cases perhaps impossible. That is no less true of more ordinary desires, like wanting a friend to get a great job, but really wanting them to get a great job that doesn't make you insecure. We will have to be content with fallible but typically good evidence. I use the term 'evidence', but it might be that these are more than evidence, that they are somehow constitutive of S's desiring x, or partly so. I don't mean to rule that out, since if it were true, these would definitely be good evidence that S has the relevant desire. To summarize, then, in figuring out whether S desires x, here are the sorts of evidence we can appeal to:

1. S pursues x, and changes how S pursues it according to how S thinks x is effectively pursued.

2. S is satisfied, content, etc., when S thinks S has gotten x (or x is true, has been

that says 'if p appears desirable (valuable, etc.), then believe that you desire that p. The differences won't matter here.

[40] Byrne notes that DES's credentials aren't quite as pristine as BEL's seem to be, because of cases of accidie (a kind of torpor). Bracket such cases for now.

realized, etc.) because they've gotten, etc., x.

3. S finds x desirable.

Let's finally return to **Objection 3** and $\mathcal{D}$. We can get all three sorts of evidence. A person can certainly pursue what they think they objectively should want, and adjust their pursuit in ways they take to be effective at getting it. They might take an ethics course, for example, and readjust the way they live their lives based on the information they receive there. People do do this, even if perhaps less often than we would like.[41] It's worth wondering how far people really can go in this direction, and I take it to be an interesting combined ethical and empirical psychological and sociological question whether any people *do* go as far as would be predicted by the general desire in accord with $\mathcal{D}$. We do have evidence that people can go pretty far in that direction, though.

Turn next to 2. A similar story seems right there. The effective altruist who learns that their money saved n lives is apt to be quite satisfied because of it. I see no difficulty that would prevent a person being satisfied when they learn or think they learn that something they should want has come to pass *because* it's what they should want.

Finally, 3 seems obviously satisfiable. You might wonder how anyone *couldn't* find what one should, objectively, want all things considered desirable. The best case I've been able to construct works as follows. Suppose p's truth would be disastrous for the world, but there is a causal chain from S's *desiring* that p to some other proposition q's truth, and q's truth would be absolutely wonderful. Is p desirable or isn't it? *Desiring* that p is surely desirable. But DES's antecedent does not have the desire's being desirable, but the potential object.[42] Indeed there might be desires like this, and it might even turn out we objectively should, all things considered, have them. That'd generate a rational requirement to have them, if we can. (It's not at all clear we can!) But the *vast majority* of cases will not be like this. We already knew DES would not cover every case, and so this particular potential failure doesn't give huge cause for concern, though it is worth flagging.

If you exhibit the behaviors and attitudes I just described in connection with 1–3, we would have some reason to think you desire in line with $\mathcal{D}$. I see no reason to think people can't, or even, really, always don't exhibit those behaviors and attitudes. It is always possible that even when they do, they don't really have the general desire those behaviors and attitudes suggest. But that is no different from many completely ordinary desires. Given both this discussion and the observation from section 1, I conclude that we have decent

[41] [Singer(2009)] has a good discussion of some students and others who did just that, giving kidneys and the like.

[42] This might remind the reader of [Kavka(1983)]'s *toxin puzzle*. In fact the basic structure of the thought experiment goes back to [Hutcheson(2002)], page 25.

reason to think that desiring in line with $\mathcal{D}$ is possible, whether or not it is actual. And if we *can* conform to it, we (subjectively, rationally) should. Doing everything we subjectively should is, of course, very difficult. We're bound to mess up from time to time. But that doesn't mean we couldn't but have messed up, nor does it mean we are not criticizable when we do. It just means we're human, and thus not doing everything we should, including having the attitudes we should. There's nothing mysterious in that.

3.6 Conclusion

There's a lot more to explore in the basic view and approach I've presented here. Even though I've tried to address the most pressing or natural objections, I know I haven't addressed all the worries one might have. But if the approach is on the right track, we have learned some interesting things. In particular, Hume likely *would* be making a rational mistake in not preferring the scratching of his finger to the destruction of the whole world. Finally, the prospect of pursuing similar projects for other attitudes that exhibit similar behavior with respect to the possibility of general attitudes of the relevant types intrigues me. I suspect that each one of those projects would differ significantly from the project with desire, since just as belief is so different from the other attitudes, so is desire.

CHAPTER 4

Deliberation, General and Particular

...then he should realize that the beauty of any one body is brother to the beauty of any other and that if he is to pursue beauty of form he'd be very foolish not to think that the beauty of all bodies is one and the same. When he grasps this, he must become a lover of all beautiful bodies, and he must think that this wild gaping after just one body is a small thing and despise it.

—Diotima of Mantinea, *Symposium*

Introduction

We form some of the attitudes we do by rational deliberation, and even more of them are subject to rational evaluation. That's not true of *every* attitude; we are sometimes simply struck with disgust or fear of a thing. Let's limit our focus to attitudes that aren't like that. How should rational deliberation about whether to have a particular attitude proceed? Contrast two views. The first is familiar from reflection on epistemology. In deciding whether to believe that anthropogenic global warming is happening, we consider the evidence for it and against it, in part by considering what else we believe, and in part by looking for other evidence. In other words, we deliberate about whether to believe the specific proposition. I'll refer to this view as DELIBERATIVE PARTICULARISM, 'particularism' for short.[1] There's a view that's opposed to that one, according to which it is not rational to focus on particular potential objects of one's attitudes. Rather, one should consider what general features warrant the attitudes at issue. Here's my official statement:

DELIBERATIVE GENERALISM. In forming attitudes of type Φ, one ought to deliberate only about which properties make objects worth bearing Φ to, and not about any particular objects themselves.

DELIBERATIVE GENERALISM, or generalism, is, I'll admit, a highly revisionary view. By that I don't just mean that it conflicts with some intuitive, pre-theoretic judgments I or most people are inclined to make, but also that I think that people more or less violate it constantly and without

[1] Please do distinguish it from [Dancy(2004)]'s unrelated view!

hesitation or compunction. Before investigating why people do that, I want to make the best possible case for the view that I can. I will argue that it follows from basic normative facts and a value-maximization picture of requirements on attitudes. But because DELIBERATIVE GENERALISM does so badly at constraining people's actual attitude formation, it really is pressing to get clear on why it should be false if it is, and where the argument for it that I present first goes wrong. So the structure of the paper is this: first, I'll try come up with the best general argument that I can for it, one that I think really is pretty compelling. In the process I'll argue for a lemma that I haven't seen argued for before, though I think it's quite plausible; I'll call it FAIRNESS FOR ATTITUDES.[2] I'll also develop some interesting objections to that argument along the way, but I'll find them wanting. In the second half of the paper, I'll try and say what I think the best possible response to the argument I gave for generalism in section 1. Since generalism is not typically formulated or defended explicitly, much of this work will aim at translating objections people have made to similar views. The two main objections I consider in section 2 are *prima facie* strong but ultimately, I think, simply encourage us to understand generalism in better, more complex ways. In other words, I will remain inclined to think generalism likely true.

4.1 The Argument

In this section, I will give what I take to be a compelling argument for DELIBERATIVE GENERALISM. In the first subsection, I will present an argument for an intermediate result, a constraint on attitudes that I take to be plausible in its own right. In the subsection after that I will address some important objections to the argument for that constraint. Then in the last subsection I will finish off the argument for generalism.

4.1.1 Being Fair to the Objects of Our Attitudes

I'll start with an analogy based in normative ethics. When we have to distribute limited resources among people who are all equally deserving, we sometimes have to make hard choices. In one standard example, we might need to decide to give a kidney to one of two people with similar life expectancies conditional on receiving the kidney, clean lifestyles, expected societal contributions, and so on. Perhaps we should decide by lottery, and perhaps not; the *fairest* thing might be to leave it to chance.[3] But when the good is guaranteed to divide up however one likes (if, for example, one has an infinitely divisible homoeomerous good), then the following seems right:

FAIRNESS FOR GOODS. In deciding how to distribute good G, when one has equal reason of the

[2] [Smith(2004)], page 269 relies on something like it in the special case of desire, but does not formulate a general principle, either for desire or for the attitudes in general.

[3] There is a large literature on this question. For the *pro* position see, among others, [Broome(1990–1991)]. For a tentative *con* position, see, among others, [Henning(2015)].

same kind to give candidates $c_1, ..., c_n$ a given quantity of G, then fairness requires giving $c_1, ..., c_n$ equal amounts of G.

When that principle applies, we can say that each c_i is owed equal treatment. If, for example, I am distributing five dollars among five people, and I have no more reason to give any more to one than any of the others, fairness requires that I give each person one dollar. I would be unfair if I didn't. This is a moral principle at root, and has an ultimately moral justification. In short, we ought to treat likes alike, even if the natural ways of doing that are sub-optimal proxies like chancy processes.

Something similar seems to be true of our attitudes. That is, it would be strange or worse if I am angry at Ivan for being intentionally cruel to my friend but not at all angry at Dmitri for being roughly as intentionally cruel to the same friend, supposing there is no further difference in the situation that rationalizes the difference in my anger. There might be moral reasons to be angry with people who do similar things that justify the anger, just as there are moral reasons to distribute goods fairly. There are, however, cases where nothing of moral significance is at stake, but where difference in attitude without difference in justifying situation seems, again, strange or worse. Suppose, for example, that Zossima is trying to figure out whether to believe that candidate X or candidate Y will win the upcoming election. Suppose further that one pundit, A, says that X will win; and another pundit, B, says that Y will win. There are two importantly different versions of this case. First, imagine *all* Zossima knows about A and B are their respective track records in predicting elections correctly: 75%. And yet Zossima, who starts out with credence .5 that A will win and .5 that B will (and no particular interest in either candidate's winning), upon hearing about the pundits' predictions assigns .75 to A's winning. Zossima would be unfair to B, and to the proposition that B wins. It seems that, since he has no reason to go with A or B, he shouldn't weigh one of their testimonies greater than the other. On the other version of the case, suppose Zossima knows more about A and B than in the previous case, for example their gender: he knows that A is a man and B a woman. But though he takes gender to be irrelevant to the reliability of a person's electoral predictions, he once again defers to A rather than B. It seems, once again, Zossima has made a mistake.

Here are the lessons I draw. First, fairness in the sense I mean now needn't be moral at all. Zossima doesn't *owe* B anything—if you like, you can imagine A and B are really just two different computer algorithms the details of which Zossima is totally unaware of—and so his decision to go with A isn't any kind of moral violation. Yet it seems like he is being irrational in assigning A more weight than B, arbitrarily, since they have exactly the same track records and he knows nothing else about them. So I take the kind of fairness I'm discussing now to principally concern rationality, and only secondarily if at all morality. The second variant on the case more plausibly *is* a moral violation; it seems to be a testimonial injustice based on Zossima's implicit bias against testimony from women.[4] For my present purposes what matters isn't that Zossima is being epistemically unjust to B, though it seems he is, but rather that he is being irrational. He takes gender to not be relevant

[4] See [Fricker(2009)] for the *locus classicus*.

to the goodness of testimony on the issue. Thus, the second version of the case should be like the first case. Thus he is being irrational. So the two lessons are: first, cases of unfairness involving our attitudes seem to be, among other things, cases of irrationality; and second, the irrationality depends at least in part on what the agent's *reasons for* having the attitude are, and not just what *explains* their attitudes being the way they are. I am interested, that is, in an agent's *motivating reasons*—the reasons *for which* they bear the attitudes they do. If Zossima gives credence .75 to X's winning in such a case *because* A does, and not because A does and A is a male and the only contradictory testimony comes from a woman, then Zossima is being unfair in my intended sense (and in other senses).

To state this all more explicitly, here's the principle I will give an argument for:

> FAIRNESS FOR ATTITUDES. For all attitude-types Φ, people S, objects o, and (possibly logically complex) properties F, if S bears Φ to o *because* o is F (in the sense of being S's motivating reason), then for all objects o', if o' is F and it is possible for S to coherently bear Φ to o' jointly with o, then S is rationally required to bear Φ to o'.

A couple points of clarification. First, I have not stated explicitly whether this rational requirement ought to be interpreted as a wide- or narrow-scope requirement. I'm not convinced the differences between them are normatively important, and so I am not sure whether it matters which we go for, but in the interests of caution, I stipulate that the principle is to be read as a wide-scope requirement. That is, the principle's logical form is '$O(\phi \rightarrow \psi)$' rather than '$\phi \rightarrow O(\psi)$'.[5] Second, I have the clause 'and it is possible for S to coherently bear Φ to o' jointly with o because there are some attitudes for which this is notoriously *not* possible. In Buridan's ass cases, for example, the lore has it that agents can see no more reason to go for (that is, to intend to go for) o_1 than for o_2, and they have to intend to go for *one* of them. I do not require people who see two equally, indeed identically appetizing treats equidistant from them to intend to go for both or neither even if it costs them their lives, and so add this qualification.(Other solutions to this situation are possible, but this is a particularly easy one.)[6] Finally, the 'because' is meant to signal that these attitudes are ones that are held for reasons; they are not just, say, feelings of hatred that strike one, whether or not one endorses them, etc.

I think FAIRNESS FOR ATTITUDES is plausible as applied to individual cases, as I argued when I discussed belief. And one way to argue for it would use induction given further cases, across a wide range of attitude-types. To do so, I would have to be careful to pick cases with no moral valence, or even a moral valence that runs against the putative rational requirement. This is possible:

[5] For early and influential moves in the debate, see [Kolodny(2005)] and [Broome(2007)].

[6] For interesting discussion along these lines, see [Ullmann-Margalit and Morgenbesser(1977)]. The original case, despite being named for Buridan, comes from al-Ghazali and involves dates, and indeed the case itself was associated with Buridan because it was re-raised as an objection to his theory of action. See [Rescher(1959)] for discussion.

Unfair hatred. Margaret hates Molly because she's Irish Catholic, but doesn't hate Ciaran, even though he, too, is Irish Catholic (say that the cause of this, which Margaret doesn't endorse, is that Ciaran looks like an old teacher of hers.)

I think Margaret is irrational, and not just because she hates people for things that don't matter, like their nationality, ethnicity, or religion (though also because of that). Even so, argument by induction would require a *lot* of examples, and more importantly would not be explanatory or particularly illuminating, so that is not the method I will pursue.

Another method might be to show that FAIRNESS FOR ATTITUDES follows from widely-held normative theories about given attitudes. This will work for some. I think, for example, that [Feldman and Conee(1985)]'s evidentialism—roughly, the view that we are rationally required to be confident in p in proportion to how much our total evidence supports p—*does* entail a version of the principle restricted to belief. Nevertheless, this route is difficult for a few reasons. First, evidentialism and, frankly, pretty much any substantive normative epistemological theory will be highly controversial, and I don't really want to appeal to what's controversial to defend what I think should not be all that controversial. Another problem is that most of the attitudes I am interested in (which is pretty much all of them) do not have associated with them worked out theories of when rationality requires having or not having them. There is, in other words, nothing like evidentialism for hatred, admiration, etc. I'd have to do a lot more constructive work to make much progress on the questions I'm concerned with here. Finally, and perhaps most importantly, I think doing so would get the justificatory relations badly wrong: one strength of evidentialism is that it coheres well with FAIRNESS FOR ATTITUDES, i.e., one should be more confident in evidentialism upon learning that it coheres well with FAIRNESS FOR ATTITUDES. So I wouldn't want to use evidentialism to in turn justfify FAIRNESS FOR ATTITUDES.

I'll pursue a different path. I think there are very general reasons for accepting FAIRNESS FOR ATTITUDES, brought out by the *Symposium* quote from Diotima with which I began. But because I don't aim at exegesis, I will say that it is inspired by what she says. Attitudes can be formed and held well or badly. They can be fitting or unfitting, apt or inapt, right or wrong, and wise or foolish, for example. A central determinant of whether an attitude is formed and held well or badly is what the *object* of that attitude is; another is what the *reasons for which* the agent forms her attitudes, if any. Forming and holding an attitude well or badly is in large part a matter of what the objects of those attitudes are and of what the reasons for those attitudes towards those objects are. For now, in fact, I will assume *nothing* determines whether S holds an attitude Φ well beyond the object o and her reasons for bearing Φ to o. This assumption will come in for heavy scrutiny later, but it is a plausible assumption nevertheless.

There is value in forming and holding an attitude well, and disvalue in forming and holding an attitude badly. In the intended sense of 'well', this is meant to be obvious: there is value in doing *anything* well.[7] Especially pertinent for my purposes is the widespread endorsement of such

[7] I say "intended sense" because, of course, there is a sense in which can commit perfidy well, say without

a thesis applied to the *attitudes*. [Hurka(1998)], for example, endorses the view that there is value in loving the good (though less value in the loved good itself), and disvalue in loving the bad; and value in hating the bad, and disvalue in hating the good.[8] More generally, at least a large part of virtue is having our attitudes well, and virtue is valuable. This idea is familiar as concerns the doxastic attitudes. There's something good *for* the person who has accurate beliefs, or beliefs that fit the evidence, or whatever your preferred doxastic good is. Agents with epistemically good beliefs secure this value for themselves no matter whether these beliefs then go on to be practically useful in securing other things of value. That's the way a family of popular stories has it, anyway.[9] The important point is just: having attitudes well is valuable for the person who has them well.

Now, consider someone, S, who bears Φ to o because o is F, e.g., someone who admires someone for their courage in speaking out against a great wrong (and who possesses no otherwise disqualifying vices). If they then go on to refuse to admire someone who is also F—who exhibits courage in speaking out against a great wrong (and who possesses no otherwise disqualifying vices), then S simply leaves value on the table that is there for the taking. That is irrational. Returning to the epigraph, that's why Diotima says that upon realizing that there are other objects of similar beauty, the lover should realize that loving a particular beautiful body alone is a "small thing". A person who loves a particular beautiful body *because* it's beautiful endorses that object as a proper object of love in virtue of its beauty. But then they ought to think that any other object that is beautiful to the same degree and, perhaps, in the same way is also a proper object of love. They ought not to leave value on the table, if they can help it.

This argument presupposes that our attitudes, such as love, are not scarce resources in the way that kidneys are. If you doubt this, you wouldn't be in the worst company.[10] Even then, though, there will be many cases where individuals haven't yet run up against the limits of what they can e.g. love where the principle will still apply. At worst, then, we would need to make the kind of exception for love that we made for intention, as I have in effect already done. For what it's worth, I think our attitudes are not in general scarce resources. I believe that my dog weighs fewer than one-hundred pounds, and that he weighs fewer than one-hundred-and-one pounds, etc. Attitudes can, of course, compete with one another for salience, and in the way they dispose us affectively; but I think it is a mistake to confuse these things with the attitudes themselves. I will drop the issue for now; it will return in a slightly different guise in a bit.

I'll now try to make this argument a little more precise and general. Here are the first two premises:

---

risk of detection. Similarly, there might be attitudes that we ought never to have, even if in *that* sense we can have them well; but in my intended sense, if we ought not to hold Φ, then we cannot hold Φ well.

[8] The view is much older, though, even if we don't trace it all the way back to Plato: it has antecedents in Brentano and the early analytic moralists. (See Hurka, fn. 1 for references.) For a more recent application, see [Srinivasan(forthcoming)]'s discussion of the value of anger.

[9] These include, but are not limited to, veritistic views like [Goldman(1987)], [Joyce(1998)], [Joyce(2009)], and [Pettigrew(2016)].

[10] See, e.g., [Freud(2010)].

VALUABLE ATTITUDES. For all attitudes Φ, there is value in bearing attitudes Φ to objects that are worth bearing Φ to.

GREATER VALUE. If there is value in bearing Φ to o, and to o', and it is coherent to bear Φ to both o and o', then there is strictly greater value in bearing Φ to *both* o and o'.

I have already argued for VALUABLE ATTITUDES. It is meant to capture the thought that there is something *good* about bearing the right attitudes, and is a generalization of the idea that there is something of value in having the right beliefs. It takes no stand on what the right attitudes are, or even whether for every attitude, there is a way to have that attitude well (e.g., anger). It is not an unassailable principle, but the same thoughts that motivate the idea that good beliefs are epistemically valuable seem to motivate this principle, too, and so I am comfortable arguing from it. One further thing: when I say that an object is "worth" bearing an attitude of a given type to, I mean an object that makes an overall positive *pro tanto* contribution to how well the attitude-holder bears the token attitude at issue.

GREATER VALUE is a minimal aggregative principle, minimal because it is neutral between the various ways the greater value might be determined. I have added a coherence constraint, since it might be that attitudes that are individually good ones to have are bad when had together. With that constraint, the principle seems trivial. It does *not* assume that, if the value of bearing Φ to o in C is x, and of bearing Φ to o' in C is y, then the value of bearing Φ to both o and o' is $x + y$; only that it is strictly greater.[11]

The next part of the argument is, in a way, stipulative: in the intended sense of 'because', when an agent S bears Φ to o because o is F, then S is rationally required to think that F-ness to make objects worth bearing attitudes of type Φ to. There are, perhaps, ways of thinking of motivating reasons that do not generate this rational requirement, but if so, they are different phenomena than the one I'm interested in.

So, from these principles, we (rationally) should think that, if we bear Φ to o because it's F, then bearing Φ to both o and o' (where o' is also F) is more valuable than just bearing Φ to o. But since the discussion is limited to attitudes that aren't scarce resources, bearing Φ to o but not to o' would then be to forego greater value for lesser value. Since it's irrational to do that, it is irrational to bear Φ to o because it's F while not bearing Φ to o'. In other words, FAIRNESS FOR ATTITUDES is true. That is the basic argument.

Before I move on, I want to consider an objection. As I have formulated it, FAIRNESS FOR ATTITUDES is a rational requirement. The objection begins by asking whether agents are rationally required to believe VALUABLE ATTITUDES (or, for that matter, GREATER VALUE, though I will focus on the former). If it's true, as I believe it is, is it not a substantive truth concerning value and the various attitudes we humans happen to have? I'm not exactly sure how substantive it is. Either way,

[11] [Kagan(1988)] has argued convincingly, I think, that normative theorists ought not simply to assume aggregative value (etc.) are additive.

though, if agents are *not* rationally required to believe it, then they are not rationally required to find the reasoning of this subsection's main argument compelling. They need not do anything irrational in not satisfying FAIRNESS FOR ATTITUDES. One thing that makes this objection so complicated is that there are various notions of 'rational' at play in philosophy at present. According to one, a belief is rationally required if an agent would be rendered incoherent in some way not to have that belief. One is probably not rationally required, in this sense, to believe VALUABLE ATTITUDES. On the other hand, there's a looser notion of 'rational' also in use, according to which, for example, it is irrational, though not incoherent, to not believe (for example) that the sky is blue, given my evidence.[12] I *do* believe that any agent with a normal adult human's evidence rationally ought to believe VALUABLE ATTITUDES (which of course doesn't mean they could or would articulate it, especially not as I have), when 'rationally' is meant in this other sense. There's something good about—something valuable in—having attitudes well, just like there's value in doing *anything* worth doing well.

All that said, you might disagree with me that there is any non-stipulative sense in which agents are rationally required to believe VALUABLE ATTITUDES. Nevertheless, there does seem to be *some* sense in which agents ought to believe it: it's true (I've argued), and fairly obvious, even. Since DELIBERATIVE GENERALISM as I stated it uses a 'should' that I have not marked with a particular flavor of normativity, if you find the objection of the previous paragraph compelling, in the statement of FAIRNESS FOR ATTITUDES, please replace 'is rationally required to' with that same 'should'.

That concludes my argument for FAIRNESS FOR ATTITUDES. I think it is also intuitively plausible, and accounts for judgments about cases well, too. But the *way* I've argued for it is important. So before I use it to argue for generalism, I want to raise some important but, I think, ultimately wrong-headed objections. That's what I'll do in the next subsection. Then in the subsection after that I'll complete the argument for generalism.

4.1.2 Local, Federal, and Global Norms

In this subsection, I'll make a tripartite distinction between kinds of attitudinal norm; mostly this distinction is familiar, but mine is a little more complicated, and so requires some elaboration. I'll then parlay this distinction into a *prima facie* objection, and then answer it. Doing so should also help clarify the sort of argument I'm formulating and examining in this paper.

It's easy to give examples of the distinction in action and see how to project from those examples, but less easy to articulate the distinction exactly, so I will begin with examples involving doxastic attitudes.

- *Local*. S's belief that p would be rationally impermissible when and because S's total evi-

[12] For roughly this distinction, see [Worsnip(forthcoming)].

dence makes p very unlikely.[13]

- *Federal*. S's high credence that p would be rationally impermissible when and because S's overall credences would be less accurate in expectation by assigning p high credence than by assigning $\neg p$ high credence or assigning no credence to p or $\neg p$.[14]

- *Global*. S's belief that p would be rationally impermissible when and because S already has a very large number of beliefs even without that belief.[15]

Let S's Φ set be the collection of all S's simultaneously held Φ-type attitudes. A *local norm* N is a norm whose recommendations take into account only the goodness or badness, fittingness or not, of the particular mental state at issue. Our evidentialist norm, e.g., recommends what it does because its point is to ensure that beliefs fit the evidence in proportion to the evidence's support. It is not concerned with the goodness or badness, or any other normative status, of the belief set as a whole.

Federal norms, on the other hand, are norms whose recommendations look to the goodness or badness, etc., of the whole Φ set—but where the goodness or badness, etc., of the whole Φ set *reduces to* to goodness or badness, etc., of *each member* of the Φ set. For example, given a particular credence function Cr and scoring rule $\mathcal{I}$, the contribution of each degree of belief in some proposition contributes to the expected inaccuracy, and nothing else does. Even if the credence set as a whole matters, its overall normative status is determined by the individual degrees of belief and their expected inaccuracy.

Finally, global norms are non-local norms whose recommendations are not determined pointwise by the goodness of the individual members, but rather depend on global features of the Φ set. Our example is a belief set's *having many members*. That some belief b is part of an individual's belief with a ton of other beliefs does not say anything about the goodness or badness, etc., of b itself.

There's a lot more to say about this distinction, but this should serve for my purposes.[16] The objection is fairly simple: my argument was federalist through and through, which both localists and globalists would take issue with. A localist about belief, say, thinks that we are rationally required to believe that p only because of how good, etc., the belief that p would be, and not because of how good, etc., the belief that p would make the overall belief set. Localist-driven anti-federalism has enjoyed a revival of late.[17] That is due in large part to examples like these:

[13] For this sort of norm, see again evidentialists like [Feldman and Conee(1985)].

[14] For this sort of norm, see epistemic consequentialists like those in footnote 8.

[15] For this sort of norm, see [Harman(1986)], page 12 and [Friedman(2017)]. Their exact norm is somewhat different: "one should not clutter one's mind with trivialities". (Friedman prefers 'junk'.) Interestingly, though that norm is naturally read globally, the norm Friedman ends up after some refinement is not: "necessarily, if p is junk for S at w, t, then S ought not to believe p at w, t". Indeed, that norm is naturally read as *local*, which is Friedman's own interpretation. Either way, my example is less plausible than either one, but illustrates the kind of norm I have in mind more effectively and exactly.

[16] Sarah Moss has work-in-progress that makes a similar distinction as applied to norms for imprecise credences.

[17] See, e.g., [Berker(2013b)], [Berker(2013a)] and [Greaves(2013)]. It goes farther back than that, though,

Converted scientist. Ivan is an atheist scientist who seeks a grant from a religious organization. The grant agency will only give grants to fund Christians' research, and Ivan knows that he wouldn't be able to fool them. Despite the fact that he's thought long and hard about the issue and knows that, on balance, the evidence strongly supports atheism over Christianity, he also knows that this will cause his belief set to be filled with more accurate beliefs than he had before that.[18]

The idea is that Ivan is rationally forbidden from believing the tenets of Christianity, at least for those reasons. Those who find such examples compelling might put forward the following explanatory hypothesis: the only genuine norms of rational attitude formation are wholly local. This is meant to capture the "separateness of the proposition" (or, better, object of our attitudes, understood generically).

That hypothesis is premature, however, and the argument for FAIRNESS FOR ATTITUDES shows how it is. By the stipulation I will examine next, local norms are the only determinants of how well a given attitude is held. Federal norms mandate having attitudes that already (are taken to) do well *by those local norms*. So, if hating someone who has gratuitously and culpably hurt a friend of mine is rationally permissible according to the prevailing local norms, then so is hating *anyone* who did the same. Such attitudes must pass the local norms' tests in order for the federal norm, FAIRNESS FOR ATTITUDES, to kick in. Such norms are perfectly compatible with the intuitive verdicts in cases like *Converted scientist*, since the Ivan who comes to believe in the tenets of Christianity for these reasons does not believe well by local norms. I submit, then, that a better explanatory hypothesis to make on the basis of such examples is this: genuine federal norms do not generate requirements to have some attitude when local norms require *not* having that attitude.[19] Of course, you might simply dismiss the intuitions in cases like *Converted scientist*; I have no argument to offer against doing that, but neither does anything I say rest on such intuitions. I just aim to demonstrate the compatibility with the argument for FAIRNESS FOR ATTITUDES with those cases and the concerns they raise. There is a larger point of interest, though, that oughtn't to be lost in the shuffle: wholesale localism cannot be motivated the way localists have tried to motivate it.

The other objection to the argument for FAIRNESS FOR ATTITUDES that I presented proceeds from the possibility that there are *global* norms. The global norm I gave above is a little silly, but that doesn't mean they all need to be. Here's one that's somewhat more plausible:

$N_{\#}$. Do not bear intense love to very many objects.

(This vague formulation will suffice for my purposes.) We must be careful about the motivation for this kind of norm, though. I stipulated before that FAIRNESS FOR ATTITUDES is only to apply to

e.g. to [Firth(1998)]. It is often pitched somewhat differently than I how I pitch it here, as about teleology, but I prefer my way of going, since it is more general: the issue is really about aggregation.

[18] See [Berker(2013a)] for discussion of this case.

[19] That still spells trouble for some varieties of federal norm, indeed for Berker's and Greaves's main targets.

attitudes that are not "scarce resources"; so, if it is *impossible* to bear intense love to many objects,[20] or if it is very costly—as a matter of use of cognitive and affective resources—to do so, then this is the kind of global norm that need not conflict with anything I've said. Other motivations would cause trouble, though. You might think a person's intense love of too many individuals, hobbies, etc., cheapens their intense love of any one of those things—it reduces its value. This could be for, e.g., *moral* reasons, according to which, for example, it is immoral (adulterous, etc.) to intensely love more than one person.[21] I could also imagine broadly aesthetic reasons. Bracket just what might motivate belief in $N_{\#}$. The idea is just that, though love would not be a scarce resource, conforming to FAIRNESS FOR ATTITUDES would not be costless, at least if the reasons motivating some token bearing of intense love are sufficiently generally applicable. Since surely when love is immoral (or cheapened, unseemly, etc.), that detracts from its value, even when the love is, say, *apt*. This is the objection from global norms: there might be *costs* to bearing attitudes, even when those attitudes are nevertheless valuable as far as local and even federal norms are concerned and are not scarce.

I would like to argue that there are no genuine global norms. That's because I *believe* there are no genuine global norms. But justifying that belief, much less persuading anyone of it, is a difficult task. The norm I just mentioned, not to intensely love too many people simultaneously, strikes me as deeply wrong-headed, but other global norms have similarly ancient pedigrees but strike me as more plausible, e.g., *everything in moderation*. I will, then, argue for this weaker claim: there are no genuine *irreducibly global* norms that tell individuals not to have too many objects of a given attitude-type. There might be other objections to my argument for FAIRNESS FOR ATTITUDES from global norms, but this is the only kind I've thought of.

I earlier gave $N_{\#}$ some sample motivations: moral ones, say concerning adultery, and aesthetic ones. I will defuse the challenge presented by $N_{\#}$ by arguing that, when it is plausible, it is so because each of its recommendations arises from some plausible *local* norm. That is, if $N_{\#}$'s relevance to a case of attitude possession indicates that having the attitude is disvaluable in that case, that will ultimately be so because there is some *local* norm that makes having that attitude bad in those circumstances. So, for example, it might be bad to intensely love someone who isn't my spouse when I'm married and have not established the marriage as open. But if so, that's because there's a (local) norm against *adulterous* intense love. That is, love is *worse*, let's suppose for the sake of argument, when that love is adulterous. Similarly, when I love a great many things intensely, it might, perhaps, be unseemly to add to that list; but then it is the unseemliness, a local reason not to love, that does the work, even when unseemliness itself depends on numbers. What the globalist who objects to my argument for FAIRNESS FOR ATTITUDES is provide some kind of justification for caring about the numbers *as such*. I cannot think of a reason for this, besides the scarcity of the attitudes themselves or the cognitive and affective demands they would place on agents, which I've

[20] For the rejection of this kind of claim, see [Jenkins(2015)].

[21] See, e.g., [Steinbock(1986)], though note that most of her discussion focuses on marriage in particular.

already excepted. My response to the objection from global norms is, then, a challenge: find me a global norm that causes trouble for my argument that does not differ in its recommendations from local norms.

That concludes my response to the localists and globalists. To repeat, I need not disagree with localists' intuitions, or with plausible local norms; and I haven't yet seen a plausible global norm whose plausibility isn't borrowed from a plausible localist norm. The argument for FAIRNESS FOR ATTITUDES still stands. In the next section, I'll show how to use that principal to argue for DELIBERATIVE GENERALISM.

4.1.3 From Fairness to Generalism

Before I present the argument, I want to make DELIBERATIVE GENERALISM more precise and explain it; that's not because the argument itself demands extra precision, but simply because I think the added precision will help situate it in a space of related philosophical theses.

I think of generalism and particularism as contrary views on what correct update policies are for Φ-type attitudes. (A less committal view is, of course, possible: both are false, and there are no true general constraints in the neighborhood.) Updating policies are very familiar from epistemology. Examples there include enumerative induction, conditionalization, and inference to the best explanation. Unfortunately update policies receive discussion almost exclusively in the context of epistemology.[22] But given that we have many more rationally evaluable attitudes, we should also have rational update policies concerning those attitudes. Deliberative generalism about Φ comes to this: agents ought only to have policies for updating their Φ-type attitudes that consider the properties objects might have that make Φ worth bearing those Φ-type attitudes to, i.e., agents ought *not* to consider whether those particular objects themselves do or don't actually have those properties.

On this view, it is a mistake, one that most people make almost all the time, to bear the relevant attitudes in what we might call a *reactive* way, that is, in the following way: some object o's F-ness (or apparent F-ness) becomes salient to an agent S, and in response, S, thinking that F-ness rationalizes S's bearing Φ to o, comes to bear Φ to o. This way of forming our attitudes is absolutely pervasive, so pervasive, as I suggested in the introduction, that it often escapes notice. To say that we should not do that is, then, *highly* revisionary. One unfortunate consequence of this pervasiveness is that even I think that any argument against doing things this way needs to be especially strong; so, even if the argument I ultimately present is good, as I think it is, it is still fair to not be moved completely to accept it. In other words, though I accept the argument of this section, the responses to the objections I develop in later sections will and should bear a great deal of the weight in generating actual conviction.

First I'll present and defend the argument, and then I'll explain what it, in conjunction with the argument for FAIRNESS FOR ATTITUDES, really amounts to. Here's the argument itself:

[22] Chapter 1 ([Drucker(forthcoming)]) is an exception.

P1. If FAIRNESS FOR ATTITUDES is true, then every agent ought to ensure they conform to its recommendations.

P2. The only way for agents like us ensure we conform to FAIRNESS FOR ATTITUDES's recommendations is by satisfying DELIBERATIVE GENERALISM.

P3. FAIRNESS FOR ATTITUDES is true.

C. So, we ought to satisfy DELIBERATIVE GENERALISM.

I have argued for P3 already, leaving P1 and P2. I don't have much to say in defense of P1. I just assume we ought to get ourselves to do what we are rationally required to do, if we can—and the more we can ensure it, the better. It's worth worrying whether we can ensure it or even get somewhat close, but in defense of P2, I'll say why we can, by saying how.

The argument for P2 is simple: any other method, that is, any method that asks whether the particular is worth bearing the attitude to, brings with it the risk of error, either through misfire or incompleteness. Over the course of a life of any length, one will *often* fail to satisfy FAIRNESS FOR ATTITUDES *unless* one satisfies DELIBERATIVE GENERALISM. Suppose S decides to bear Φ to o because o is F, and then investigates o' to determine whether *it* is F, too, or even just reacts to o''s *apparent* F-ness. For pretty much any F, S could be wrong that o is in fact F. That's so even when 'F' is a predicate like 'is believed by S to be G', for some G.[23] That is, investigating o' *in particular* brings with it a kind of epistemic risk. This epistemic risk is eliminated once one decides that F things are worth bearing Φ to (as, recall, one has when one bears Φ to o *because* o is F), and goes on to bear Φ to *whatever* is F. So, in particular, you might decide that racists are worth despising, and thereby come to *despise racists*, rather than to investigate some set of people with the aim of determining whether they are in fact racists. Bearing Φ to what cannot be known infallibly to be F brings with it risks, and these accumulate over time. There are no such attendant risks if one deliberates generally and forms their attitudes that way by, e.g., despising whoever is racist. Anything but DELIBERATIVE GENERALISM is a risky way to rationally deliberate about attitude formation, at least insofar as one aims to conform to FAIRNESS FOR ATTITUDES.

That, incidentally, is the mechanism that makes ensuring conformity to FAIRNESS FOR ATTITUDES possible: general attitudes. Examples include sympathizing with anyone who can't pay their hospital bills, hating all racists, and being jealous of everyone who got to go to Woodstock. Because they're so everyday, I assume they're possible. I also will take it that if Grushenka got to go to Woodstock, then Grushenka is a *counterexample* to my claim to be jealous of everyone who got to go to Woodstock if I'm not jealous of Grushenka (at least for that). That is, if I'm jealous of *everyone* who got to go to Woodstock, then I must be jealous of Grushenka, too. This will be, perhaps, a little more controversial. There are worries about the way general attitudes like those will lead to weird consequences for the ascription of belief.[24] In other work, I argue that belief (and,

[23] See [Williamson(2000)] for this general thought.

[24] See, e.g., [Blumberg and Holguín(forthcoming)].

I think, intention) work differently from these other attitude-types, in that forming general beliefs of the right kind is much harder.[25] I will take those arguments for granted, then, and accordingly restrict generalism: I mean for it only apply to attitudes that are not like belief (and perhaps intention) in this way. I have no definite list to offer, but generalism would still be true of most kinds of attitudes we have. For attitudes like jealousy, etc., I think it follows *logically* that Grushenka would be a counterexample to the general jealousy attribution. I will assume that going forward.

I can conform, then, to FAIRNESS FOR ATTITUDES by bearing Φ to *every* F thing, when I take objects' F-ness to make those objects worth bearing Φ to (and with the exception of belief and perhaps other attitude-types). And indeed, this does seem like the best way; anything else opens one up to risk, that is, opens one up to bearing attitudes to objects that—by one's own lights—aren't worth bearing the attitudes to. Generalism is the only way of guaranteeing conformity to FAIRNESS FOR ATTITUDES in a way that respects one's opinions about what is worth bearing attitudes on what bases. That is, generalism is true.

I have argued from FAIRNESS FOR ATTITUDES to generalism. The argument for FAIRNESS FOR ATTITUDES was, essentially, that other ways to have attitudes would leave value on the table. So, the reason to conform to generalism is this: it maximizes value *in expectation*. Conformity to FAIRNESS FOR ATTITUDES maximizes value itself, and conformity to DELIBERATIVE GENERALISM is a way of guaranteeing conformity to FAIRNESS FOR VALUE. That is why generalism is true—if it is. In the next section, though, I will investigate whether my argument for generalism goes wrong, and if so, where.

4.2 Anti-Generalism

Few if any people are thoroughgoing generalists, I conjecture; certainly I am not in the way I *actually* form my attitudes. If there is a strong case to be made for generalism, it becomes more pressing to understand why so few people conform to it. In particular I'd like to see what the strongest objections to it are; even better would be an objection that responds specifically to the argument I presented in section 1. That is what I'll try to produce in this section, and I'll see whether and if so how the generalist can respond.

I'll focus my attention on two potential objections, one from the value of experimentation, the other from (what I'll call) alienation. Since my argument was, in essence, a dominance argument that appeals to value, the proper way to object to it would be to show that there are trade-offs that I didn't mark before. I will focus on two challenging candidates: experimentation and alienation. They both concern valuable things that the generalist is bound, either by the letter or spirit of her position, to miss out on. Since, again, that entails that generalism involves a trade-off in values, that would be a compelling response to the argument for it.

[25] Specifically, chapter 1 ([Drucker(forthcoming)], section 2.

4.2.1 Experimentation

The first objection is that the generalist does not easily find a place for experimentation. It goes like this: when I come to decide that F-ness suffices to make objects worth bearing Φ to, then I ought to bear Φ to every F thing. The thoroughgoing generalist bears their attitudes in an uncompromising way that rules out experimentation. But there is value in experimentation. So, generalism is false.[26]

This objection involves two claims. First, that experimentation of a particular sort is valuable (in the right way, I'd add). Second, the generalist cannot endorse this kind of experimentation as rational. So, what kind of experimentation do I have in mind? Here's the kind of case.

> *Miss Lonelyhearts*.[27] Miss Lonelyhearts has never been in love, and indeed thinks it's impossible to know *whom* to love without learning from *being* in love. He meets someone, A, who elicits the right physiological and affective responses, and so decides to throw himself into it, since he thinks it will be very important for his life to know whom to love.

Miss Lonelyhearts loves rationally, I claim; he undertakes an adventure in order to teach himself important things. More generally, when we don't know what features warrant bearing Φ to but where we reasonably think that bearing Φ to o will give us a good answer to that question, then it is rational to bear Φ to o. Not every sort of experimentation is like that, and perhaps not every sort of experimentation that is problematic for generalism, but that is *one sort* of experimentation that I take to be rational.

Why must generalism rule out cases like that? It depends on what stage we're focusing on. Interestingly, it does *not* rule out his coming to love A in the first place. He could come to love A by loving *whoever is suitable to teach me whom to love* (let's suppose this is psychologically possible, for the moment), where only one person, A, answers to that description. But it does seem to rule out the learning process once he's already come to love A. Return to the statement of DELIBERATIVE GENERALISM: it says that in deliberating whether to have a particular attitude of type Φ, we should only consider general features, and not whether some particular object is worth bearing Φ to because of those features. But Miss Lonelyhearts experiments in order to *discover* those features. He *needs* to focus on whether A is worth loving because of the particular features A has. Simply thinking about general features would not do the work that focusing on A *in particular* can do. So, the kind of experimentation that motivates Miss Lonelyhearts' loving in the first place is ruled out by generalism itself. Experimentation is valuable, though, so generalism is not cost-free. The argument for it therefore fails (or so the objection claims).

The generalist could respond by claiming that the experimentation in *Miss Lonelyhearts* isn't a part of deliberation. The problem with this response is that it surely is: Miss Lonelyhearts is

[26] I am inspired in this objection by [Nussbaum(1979)], who, in a wonderful study of the *Symposium*, argues that Alcibiades' response to Diotima includes the idea that there are some things we can only learn from loving particulars. The point generalizes, but must be put carefully in order to extract a problem for generalism rather than just for Plato.

[27] With apologies to West.

deliberating about what kind of people to bear Φ to, and, even if this is a quite extended process, it does not seem different in kind from ordinary deliberative processes. Another response, then, would be to modify DELIBERATIVE GENERALISM. They could change it to the following, for example:

> DELIBERATIVE GENERALISM 2.0. In rationally deliberating about whether to have a particular attitude of type Φ to objects $o_1, ..., o_n$ that we do not presently bear Φ to, we should consider what general features make objects worth bearing Φ to, but not whether any o_i is worth bearing an attitude to because it has those features.

This does avoid the problem, but I think at the cost of betraying the spirit of generalism. The trouble is that for the generalist, to deliberate about whether to bear Φ to *some* o because it is F is to deliberate about whether to bear Φ to *all* such o. And don't forget, part of the experimentation is to determine whether *A*, among others, is worth loving. So the restriction to objects to which we do not presently bear Φ seems *ad hoc* and, as I said, against the general spirit of the view.

I think experimentation does pose a problem for generalism. The problem can be put summarily like this: we deliberate about what to bear Φ to better *by focusing on particulars*. I want to emphasize that this objection relies on the empirical premise that we do sometimes deliberate better in that way, or anyway could.[28] That said, though, I don't think the objection goes very deep. Distinguish between *stages* of deliberation. Although I would like to not commit myself to any analysis of deliberation's structure, here's one kind of division you might make: determining what reasons are relevant to whether to bear Φ to some or each member of some class; determining what the relevant class of potential objects is; determining whether the reasons apply to some or each member of that class; and applying the reasons to form the attitude to the relevant objects. Then the deliberative generalist can be seen as saying something like this: *once the relevant reasons have been determined*, we ought not to focus on the particulars themselves in deliberating about whether to bear the attitude to those objects. Unlike previous modifications, this modification is not *ad hoc*, and it coheres well with the general spirit of generalism. Overall, I don't think much of interest or value in the original proposal is lost by switching to this one. (That said, much more work should go into making this division of deliberation into stages empirically and philosophically satisfying. What I've done is just a rough first cut.) So, for the remainder of the paper, please understand 'generalism' to refer to that claim. The next objection is meant to target even that version of the position.

4.2.2 Alienation

I've been suggesting (though not positively asserting) that generalism is a Platonic doctrine; so it is interesting that one of Plato's best and often most sympathetic exegetes, Gregory Vlastos, inaugurated the modern resistance to generalism. Here is what Vlastos says:

[28] For just one example of a philosopher who is congenial to this possibility, see [Johnston(2001)].

> We are to love the persons so far, and only insofar, as they are good and beautiful. Now since all too few human beings are masterworks of excellence, and not even the best of those we have the chance to love are wholly free of streaks of the ugly, the mean, the commonplace, the ridiculous, if our love for them is to be only for their virtue and beauty, the individual, in the uniqueness and integrity of his or her individuality, will never be the object of our love. This seems to me the cardinal flaw in Plato's theory. It does not provide for love of whole persons, but only for love of that abstract version of persons which consists of the complex of their best qualities. This is the reason why personal affection ranks so low in Plato's *scala amoris*. ... The high climactic moment of fulfillment—the peak achievement for which all lesser loves are to be "used as steps"—is the one farthest removed from affection for concrete beings.[29]

This is a pregnant quote; there are a lot of different worries expressed in it, some of which apply to Plato but not every sort of generalist, and some of which (it seems to me) really apply to neither. The generalist as such need not say that we ought only to love people for their best qualities, let alone only for their virtue and beauty. Nor does it say we ought not to love the people themselves, rather than complexes of properties. It is, as I said, a thesis only about proper *deliberation*, and so it is generalism in its deliberative guise that must be targeted.

Yet there is an important worry here for generalism. The idea at the end is that the way the generalist loves is somehow incompatible with affection for concrete beings. If "concrete" is read as opposed to "abstract", then the objection is not on target. But there *do* seem to be aspects of attitudes, attitudes like love, that are determined by how they were formed. To see this, suppose I hate whoever supports taking away health care from children in poverty, and I came to that hatred in the way the generalist recommends. Suppose also that I know that *Fyodor* supports taking away health care from children in poverty; we've debated this before, and I reached my limit at some point and started hating him for that reason. The two attitudes have the same reasons, and let's just stipulate are held with the same "strength" (where "strength" of hatred is roughly like strength of desire or confidence). But my animosity toward Fyodor is *personal*, because of the way it was formed. My hatred of all the other people who support taking away health care from children in poverty is, on the other hand, *impersonal*. This difference is registered by the difference in how they feel, but, I think, is not exhausted by it. There's a difference between the attitudes even when, for example, I'm depressed and not feeling much in the way of affect about anything at all. This, then, is the worry concerning alienation: there is something valuable in having *personal* rather than *impersonal* attitudes. But the generalist requires us to have impersonal attitudes. So, there's a trade-off in conforming to generalism.

It is difficult to say what alienation in general comes to. It's a fraught concept, and I have no analysis to offer of it. Famous examples have come down to us, such as [Williams(1981b)]'s man who, in saving his drowning wife, thinks to himself that it is permissible in such circumstances as his to save one's wife; another is [Railton(1984)]'s examples of spouses and friends who come to treat those people as morality (and decency) requires by employing in deliberation the bloodless

[29] [Vlastos(1973)], page 31. [Nussbaum(1979)] pulls the same quote.

formulations of consequentialism and deontology. The lesson I take from these examples, and in particular Railton's, is that the etiology of an attitude can matter for the nature of the attitude in question, beyond where it falls in some classification into, say, fear, admiration, hatred, etc. There are some attitudes, in particular the ones we bear to humans, where those differences matter.

The other lesson I take from such examples, though, is that it is very hard to say with any precision *how* the alienated deliberative process changes the product if it's not by affecting the strength or classification of the attitude. Return to the Fyodor example. I'm tempted to say that my hatred of Fyodor is more *about* Fyodor than it would have been had he just been one among the many people I hate for supporting taking away health care from children in poverty. But that, too, is hard to substantiate: not much about Fyodor besides that fact went into my coming to hate him, so in that sense, my attitude to him is no more about him than any of the others. Still, there is at least this difference: my hatred is focused on him. Even if he were to stop supporting that sort of thing, or never to have supported it, I would still hate him. But it's hard to see why that counterfactual could be important, much less ground a difference in the value of the attitudes.

Railton puts the problem with alienated people this way: "there would seem to be an estrangement between their affections and their rational, deliberative selves; an abstract and universalizing point of view mediates their responses to others and to their own sentiments".[30] Why is it more valuable, if and when it is, to have attitudes that are not mediated in that way? Railton doesn't say. Perhaps the only thing to be said here is that we don't *want* either to have or to be the objects of such alienated and impersonal attitudes. That much seems to me to be right, and perhaps that is enough to motivate the objection from alienation. When we form attitudes as the generalist would have us, we cause ourselves to miss out on something we and the objects of our attitudes might and often do want. And perhaps this desire is relatively basic and cannot be justified. In the remainder of this section, I'll sketch a hypothesis about what I think lies behind the desires, and then I'll give a possible response on behalf of the generalist.

Some attitudes can be held maximally well even when the person who has them does not care about their objects.[31] Belief is like that: if I have great evidence for believing that the number of hairs on A's head is even, then even if I don't care about that fact, I believe it maximally well when I believe it on the basis of that evidence. Other attitudes aren't like that. I might have the best evidence in the world that the person behind the curtain in front of me is a wonderful human. But my love of this person is relatively hollow unless my care of this person motivates the love. That does *not* mean the phenomenology needs to be especially strong. A parent might have the strongest, best kind of love for a child even at moments when they are not feeling particularly warm toward her, just because the love is still there and motivated by an intense concern. For some attitudes, the

[30] Page 137.

[31] Note that some philosophers use 'care', 'caring', etc., that makes it sound like a pro-attitude. (See [Frankfurt(1999)], for example.) I use it more broadly than that, like when people say 'Of course I care that he did something so horrible; I just don't know what to do about it right now." I have no analysis of caring to offer, though I don't mean to use it any differently than it is used in ordinary language.

objection goes, care matters *as a motivation*. I say "as a motivation" because I might in fact care about the person behind the curtain, under another guise, but insofar as that care doesn't motivate my love, the love isn't good as it could be. The main claim this objection makes, then, is this: for some attitudes, love and hatred prominent among them, that care figure into their motivation in the right way makes a difference to how good that attitude is. The care is itself not a *reason* to love or hate them. The reasons are just, say, that the person is a wonderful human, or that they support taking away health care from children in poverty. It is just a background fact that makes the quality with which the relevant attitudes are held different. Remember in my argument for FAIRNESS FOR ATTITUDES that I assumed that the only thing that made for how well an attitude is held is what the reasons for it are and what the object is. This objection challenges that assumption: *how* it was formed, independently of the reasons for it, can make a difference, too.

I don't know which attitudes this condition is true of, or of which it isn't, though some are much more plausible than others. The possibility that it holds of *any* constitutes an objection to generalism, at least on the condition that, in some cases, for care to motivate an attitude toward o, that attitude must be formed by focusing on the individual o. I can hate Fyodor in virtue of hating whoever is extremely malicious, but that hatred for Fyodor will differ from the kind that sees him working tirelessly to take away health care from children, say by lobbying fence-sitting Congresspeople in particular ways. This is not a question of whether people *can* hate Fyodor in virtue of hating whoever is extremely malicious; I am assuming we can. But perhaps there might be extra value in hating someone because you saw their wrongdoing and care about it.

I agree that, at least for some of our attitudes, it is better to have them in an unalienated way. Love that's not motivated by care, if it's still love, is colder than it would otherwise be. I think the generalist can admit that, and still resist the objection, since it relies on the this assumption: general attitudes cannot be motivated by care. This assumption is somewhat intuitive, but clearly false: there is no in-principle difficulty with general attitudes' being motivated by care. If I'm hurt in a particular way, I can hate anyone who hurts people in that way; that will be a general attitude very clearly motivated by care. That sort of attitude is not alienated, at least not in any way I can see. It is true that there is a chilly, detached kind of attitude formation it might look like the generalist recommends, but there is nothing in generalism that requires it, and for some attitudes the generalist is free to positively reject it on other grounds. The objection fails, I think.

As I said, though, something about the assumption feels true. I think there is indeed something right about it, something that I think has consequences for how we apply generalism, even if not for whether generalism is true. Take two descriptions D_1 and D_2, and suppose that after deliberating we come to think that it's worth bearing to Φ to anything that satisfies either D_1 or D_2 (or both). There might nevertheless be differences between D_1 and D_2. Suppose, for example, D_1 is 'person who wronged me in some way today', while D_2 is 'person who defrauded me today'. Then even if D_1 and D_2 are coextensive (and perhaps even if I *knew* they were), hating whoever satisfies D_2 might naturally involve more care than hating whoever satisfies D_1. That would mean that forming

my hatred on the basis of D_2 would be better than forming it on the basis of D_1, at least given our earlier assumptions.

Here's why all this matters. When one first considers DELIBERATIVE GENERALISM, it's natural to think that the kinds of descriptions it would commend to our deliberative attention would be things like 'objects that are worth bearing Φ to'. That's what I assumed, anyway. And it's true that such descriptions don't get the blood boiling. But some general descriptions *do*. This might simply be a contingent fact about us humans, but even so, there is an important theoretical point here. If there are attitudes whose quality depends in part on the other attitudes (like caring) that motivate the formation of tokens of those attitudes, then there will be crucial trade-offs between, on the one hand, descriptions that guarantee that the objects of our attitudes are fitting, apt, etc., objects for those attitudes, and on the other descriptions that offer less of a guarantee but allow us to have the attitude better because motivated by care. Generalism would be a massively revisionary doctrine if it recommended that we only ever bear Φ to attitudes *via* the description 'objects that are worth bearing Φ to'. But, perhaps because of contingent limitations, though perhaps not, generalism recommends nothing of the sort. It is still revisionary, but not revolutionary. With at least some of our attitudes, even the committed generalist is left with a kind of optimization problem.[32] For such attitudes, there are no *a priori* best descriptions to use. That might very well be an important lesson that the objection from alienation has to teach us.

I can't say that worries about experimentation and alienation are the only reasons why one might resist generalism. Perhaps the whole project I've been engaged with here overintellectualizes our attitudes in objectionable ways.[33] I would merely point people who worry about this to the restriction with which I began, to attitudes that we *do* form on the basis of rational deliberation. At worst, I think, I might have to change some of my examples. In absence of other, more compelling reasons to resist, I think there is good reason to incline toward generalism.

[32] And perhaps one that the generalist ought not attend to too frequently! It's likely that many attitudes are formed better when not formed as the solution to any kind of optimization problem!

[33] [Frankfurt(2006)], e.g., thinks love is not formed for any reasons.

CHAPTER 5

A Rightness-Based Theory of Communicative Propriety

Introduction

We can express a variety of mental states with language; belief is one, but so are desire, hatred, contempt, and joy. Expressions of belief—assertion, roughly—can be inappropriate, and philosophers have recently exercised considerable ingenuity and subtlety in determining when they are. They have paid significantly less attention to the conditions under which we may properly express any of these other mental states. That is not, I think, because it is a less worthy topic. So what might the reason be?

Perhaps they tacitly accept a picture like the following. When I express my belief that p, it is, at least in the normal case, or at least when we express those beliefs through *assertion* in particular, because I hope to give *you* the belief that p.[1] But when I express my anger or my joy, I don't mean to transmit that anger or joy to you, at least not in the normal case. On this view, norms governing the expression of mental states affect only those speech acts we use to infect others with our mental states. I doubt that's true, but that's a dispute for another day. The real problem with that sort of view is just that we *do* often aim to transmit our mental states through our utterances. We attempt to communicate our feelings all the time, and I do mean 'communicate' literally. (I will always mean it this way in what follows.) When some news comes out that I find disappointing, my 'it's a shame, isn't it?' does not simply assert that it's a bad thing, nor does it just *express* my disappointment in the way that a downcast face might; I hope to get you to be disappointed by the news, too, by *communicating* my disappointment *to you*. Or, more worryingly, demagogues frequently communicate their hatred to others.[2] There's a large number of linguistic devices by

[1] [Heck, Jr.(2002)] calls this the "Naïve Conception of Communication", [Egan(2007)] calls it the "belief transfer model", and [Moss(2012)] calls it the "Package Delivery Model". It is ultimately an in some ways rough distillation of [Stalnaker(1978)]'s views.

[2] For the thought about communicating hatred and other feelings, see [Langton(2012)]. For the idea that we communicate *desire* in conversation, see [Pettit and Smith(1996)].

means of which we can do these things,[3] but that doesn't mean that the use of those devices is less norm-governed than the communication of belief *via* plain-old assertion is.

To fix some jargon, where Φ is an attitude[4] (e.g., being disappointed at some particular news, or hatred of some group, etc.s), call a speech act consisting of the use (utterance, inscription, etc.) of a sentence σ a Φ-*assertion* iff the speaker publicly purports to intend to communicate Φ by expressing Φ with σ.[5] Is there anything fully general we might say about when Φ-asserting is appropriate, and when not? I'm interested in two sorts of view that accept that there is some general answer. First:

FITTINGNESS NORM of Φ-ASSERTION. For all speakers S and mental states Φ, S must: Φ-assert only if Φ is *fully fitting* for S.

Though 'fitting' is a modish term, others use different terms to express what I mean to express, such as 'correct'[6] or 'apt'.[7] I'll use all these interchangeably to express the condition that one meets when one is angry *because* someone just kicked your dog for no reason, or sad *because* your dog died, or happy *because* you have a new adorable puppy full of vim and vigor.

Fittingness theories are speaker-directed, in that for them the propriety of a Φ-assertion depends on the speaker but not the hearer. It's natural to think of the various norms of assertion proposed in the literature as instances of the fittingness norm of assertion, spelling out more precisely what the theorist thinks a believer has to be like in order for their belief to be fully fitting. At least, it's natural to think of them that way if those norms are fundamental rather than derived. So, for example, someone who thinks that S's belief that p is fully fitting just in case S knows that p and accepts FITTINGNESS NORM OF BELIEF-ASSERTION will accept the knowledge norm of assertion.[8] Others who think all that's needed for full fittingness is truth will accept a truth norm of assertion.[9] And other ways of spelling out fittingness for beliefs will deliver still other norms.[10] But as I said, those norms might be *derived* from a more basic sort of norm.[11]

That said, here's the second sort of theory:

RIGHTNESS NORM of Φ-ASSERTION. For all speakers S, hearers S', and mental states Φ, S must: Φ-assert to S' only if it is right (i.e., not wrong) for S to make S' have Φ by Φ-asserting.

[3] [Potts(2007)] provides many examples, on which more later.

[4] I use 'attitude' and 'mental state' interchangeably in this paper, though not necessarily because I think they are the same thing.

[5] I say that this is *jargon* because I don't want to take a side on whether assertion has to be individuated by its constitutive norms, as [Williamson(2000)] does, as opposed to what I just did. (For doubts about the constitutive view, see [Maitra(2011)] and [Cappelen(2011)].)

[6] See, e.g., [Gibbard(2005)] and [Engel(2013)].

[7] See, famously, [Gibbard(1990)], but also, e.g., [Srinivasan(forthcoming)].

[8] See, among many others, [Unger(1975)], [Williamson(2000)], and [Turri(2016)].

[9] See, e.g., [Weiner(2005)].

[10] E.g., [Lackey(2007)], [Douven(2006)], and [Maitra and Weatherson(2010)].

[11] Thus, I emphasize: even though I will ultimately defend a rightness norm, even if, say, the Knowledge Norm is naturally thought of as a fittingness norm, I need not (and in this paper, do not) reject it. In fact, I derive it.

Rightness theories have it rather that both the speaker and the hearer matter. I have not seen them much discussed in the literature on norms of assertion, which is somewhat surprising, since I think of them as having something like a default status. One of the aims of this paper is simply to provide the best version of the RIGHTNESS NORM that I can. The other aim is to argue for it, especially as against the FITTINGNESS NORM. One interesting fact that will emerge from the discussion is that belief-assertion (hereafter, 'assertion') is unique in a way that distorts our overall picture of communicative propriety if we focus on it too narrowly.

Though I mean to argue for a rightness theory as against the fittingness theory, it's important to understand why the fittingness theory should have proved so attractive. In the first section, I'll provide one argument for it; then I'll give a recipe for generating counterexamples to FITTINGNESS NORM OF Φ-ASSERTION that reveals the flaw in that argument. The counterexamples will also provide motivation for looking at rightness theories and provide some desiderata, and in section 2 I'll provide a framework that naturally delivers RIGHTNESS NORM OF Φ-ASSERTION. In section 3, I'll consider another argument for fittingness theories, and show how the work done in section 2 allows the rightness theorist to answer that argument. In section 4, I'll formulate a sufficient condition for the lack of fittingness constraints on a Φ-assertion, and I'll address an objection. Section 5 concludes.

5.1 Unfitting but Appropriate Φ-Assertions

In this section, I'll begin motivating the RIGHTNESS NORM against the FITTINGNESS NORM. I will first introduce counterexamples to the latter. I'll then diagnose the counterexamples: because they don't contravene the RIGHTNESS NORM, the Φ-assertions I display are not improper. But first I'd like to say why the FITTINGNESS NORM should be tempting.

We ought not to have unfitting attitudes, you might think; and if we ought not to have them ourselves, then, at least in the typical case, we ought not to give them to others. So, if I ought not to believe that p, say because my evidence for it is slim, then I ought not to give *you* the belief that p. On this view, standards for *having* an attitude translate readily into standards for *communicating* that attitude. Fittingness norms tell us when we ought not to have a given attitude. So fittingness norms will tell us when we ought not to attempt to *transmit* that attitude.

This argument, I think, trades on an equivocation involving 'ought'.[12] This section will be devoted to expanding on and defending that claim.

I'll begin with amusement. Jokes are, among other things, purported attempts to communicate amusement or mirth. A joke, especially one that the teller themselves seems to find genuinely funny, purports to communicate the joke-teller's amusement. It is an amusement-assertion, using my jargon. Now, I've had friends who are themselves hilarious but who from time to time tell bad

[12] [Brown(2011)] and [Hawthorne et al.(2016)Hawthorne, Rothschild, and Spectre] also argue against this argument, but not in the same way I do.

jokes. When I say that they're hilarious, I mean that they have some special ability to bring a room to tears with what I would say is objectively bad material; somehow, they pull it off. The joke might be your run-of-the-mill toilet humor, or it might be a *bon mot* that actually makes no sense, or it might be an overused Internet meme. Suppose the joke-teller (call him Bryan) says:

(1) Where's the beef?

Also suppose that there is nothing that *really* makes (1) funny. To make the case watertight, imagine that Bryan does indeed usually have a funny way of delivering even bad jokes, in fact he's so reliable that we've all formed enough of an expectation that he *will* be amusing that we take him to be amusing even when he isn't. Suppose this time nothing in Bryan's manner of delivering (1) is especially amusing. Yet we are amused.

According to the FITTINGNESS NORM OF AMUSEMENT-ASSERTION, Bryan's (1) was improper somehow. That's because it's both an amusement-assertion and it wasn't *actually* funny. Bryan's amusement at his own joke was not fully fitting, indeed not fitting at all. But it amused us, and it did so predictably. My first claim about (1) is that it was not wrong for Bryan to make the joke. Though there was, perhaps, *something* wrong with it, since it was unfunny (and unfunniness is a defect in a joke), Bryan did not do anything we would or should criticize him for doing. After all, we were amused, and it's pleasant and harmless to be amused, even when the objects of our amusement are unfitting as objects of amusement. Playing to the crowd might be artistically bankrupt, but it's not inappropriate in the way, say, asserting what one does not have evidence for is inappropriate. Children's authors who write books about flatulence needn't be making any error like that. None of these people violate any conversational norms, even if their contributions are aesthetically flawed.

There are other examples like this for other mental states, like joy. I take it that taking delight in something mundane's happening isn't typically fitting. Nevertheless, I don't think there's anything wrong with getting someone else to share your delight in something mundane's happening. This can, of course, be difficult. But if one is a prisoner that has little to delight *in*, then it's not inappropriate for one to get the other to share their excitement that the same (bad) food is being served once again at the regular time. Even though the joy is itself unfitting, it's not inappropriate to communicate it, or so I think. This is another class of counterexamples to FITTINGNESS NORM OF Φ-ASSERTION.

It's not difficult to extract a general recipe. Where Φ is an enjoyable or otherwise intrinsically good mental state that one can expect one's interlocutors to reasonably and appropriately want, and having Φ in the circumstances is unfitting, it is often not inappropriate to make one's interlocutor have Φ. We can usually (but not always) assume that our interlocutors would want to be amused, or delighted. This recipe isn't infallible. Spreading a belief that one has no evidence for, even if it is a pleasant belief to both you and your interlocutor, can be inappropriate. For now, though, I only need that there are many straightforward counterexamples.

What explains why this recipe generally generates counterexamples? One possibility is that it is not inappropriate for someone to (purport to) communicate a pleasing attitude. That is, of course,

false, as we've just seen. And it's very importantly false: flattery is a pernicious wrong. I think, rather, that being amused by (1) is in some ways significantly good and in no ways significantly bad. That goes for both the amusing people and the people who are amused. Similarly, being delighted is often in some ways significantly good and in no ways significantly bad. Of course, delighting people, exciting them, etc., can be a cynical marketing ploy. But when it's not, it is perfectly fine to do so. More generally, in both cases, though the underlying mental state Φ being communicated is itself unfitting, having that mental state is significantly good without being bad—they are harmless goods.

We're now in a position to see what was wrong with the argument with which I began this section. In *some* sense, all of our attitudes ought to be fitting. That is, if I could have Φ fittingly or unfittingly, I should prefer to have it fittingly, *ceteris paribus*. But with some attitudes, that the attitude is unfitting is a very weak consideration against having that attitude. In particular, there are some attitudes where the goodness of *having* the attitude far exceeds the badness of having the attitude unfittingly, and *where there need be no suitable alternative that would have the same effect*. Amusement and joy are two of these attitudes. It's not wrong to amuse someone, even if the object of their amusement is unfitting, at least so long as the unfittingness itself isn't horribly immoral, say. So, it's not wrong to communicate your amusement in even a bad joke. Though our attitudes ought to be fitting, *ceteris paribus*, not everything is equal: if I won't have the intrinsically good attitude at all if I don't have it unfittingly, then I ought to have it unfittingly. That is why the argument for the FITTINGNESS NORM OF Φ-ASSERTION equivocates.

FITTINGNESS NORM OF AMUSEMENT-ASSERTION says that (1) is inappropriate or a mistake of some kind, and to that extent that theory itself is mistaken.[13] That is reason to think that the schema of which it is an instance, FITTINGNESS NORM OF Φ-ASSERTION is wrong. RIGHTNESS NORM OF AMUSEMENT-ASSERTION does not have this consequence, and that is some reason to like it.

I have introduced an argument for the fittingness norm, and I have tried to show that the argument fails; in fact these norms have counterexamples. This makes it a good idea to see how a rightness theory should work. I've also established a couple of desiderata for such a theory. It should allow that communicating amusement is typically perfectly fine; this should have something to do with the goodness of amusement for the addressee; and the account should be contrastive. In the next section, I'll present a rightness theory that satisfies those constraints.

5.2 A Framework for Communicative Rightness

It's hard to defend the RIGHTNESS NORM OF Φ-ASSERTION without saying when it is right to make someone have a mental state. That is what I'll attempt to do in this section.

[13] Nor is the amusement-assertion *unsuccessful*, in [Mehta(2016)]'s sense; it amuses us, and Bryan knew that it would.

The fully spelled-out RIGHTNESS NORM will have two features, both of which I've already highlighted. First, it will be *contrastive*: the rightness of a given Φ-assertion will depend on salient alternative Φ-assertions the speaker might have made. Second, it will be *goal-based*: the rightness of a given Φ-assertion will depend on what goals it helps and which it hinders.[14]

At any given stage of the conversation, each interlocutor will have a number of interests, dimensions along which her life can be made to go better or worse. Conversation is principally a tool that people use ideally to make one another's lives go better, and less ideally to make their own lives go better. I will assume that communicative norms, being norms, target the ideal case. Given this, whether a Φ-assertion is right will depend on how well it advances the interlocutors' interests. I want to be upfront about the choice I've just made: this is a relatively consequentialist perspective on communicative norms. I can no more defend that choice against some standard deontological objections than I can defend consequentialism itself against them, certainly not in this paper. Nevertheless I think it is valuable to work out a minimally consequentialist view of communicative norms, to see how much of the standard apparently deontological norms we can recover by thinking in this way.

The interlocutors in any given conversation will each have a set of *actual* interests, and a set of *apparent* interests. Often an interest will appear on both lists—for example, not to be in excruciating pain—but sometimes the lists will diverge. Thus we need to distinguish between *subjective* rightness and *objective* rightness.[15] In this paper I will pretend like the lists are always the same, in the interests of simplicity. Thus I will limit myself to what are both plausibly actual and apparent interests. I will also limit myself to two-person conversations, again to keep things simple. That will limit the degree to which we must distinguish between objective and subjective rightness, but it does not extinguish the need entirely.

As I said, I'm interested in cases where the interests are *mutual*. Prior to entering into a conversation, interlocutors will have many interests in common, at least if they're normal and minimally decent people. Some interests will *become* mutual because one of the parties indicates that they have that interest. So, for example, I might strike up a conversation with you hoping to get some information about when the museum opens, and when I ask, if you don't hate me or have some reason for me not to go to the museum, it'll be part of our mutual interests that I learn when the museum opens. Some conversations are more obviously adversarial than that. But even then we will assume a store of mutual interests, at least where explicitly divergent interests aren't concerned.

So, suppose X and Y are in a conversation, and X and Y have mutual interests $I_1, ..., I_n$. Suppose finally that these facts are publicly known by both X and Y. The last analytical tool I'd like to introduce is the *salient alternative set*. At any given stage of the conversation, given any possible Φ-assertion X or Y thinks about making by means of a sentence σ, there is a salient alternative set $\mathcal{ALT}(\sigma)$: a set of alternative conversational contributions X or Y might make instead of the actual

[14] I take both of these features to have their origins in [Grice(1975)], though he puts them to quite different uses.

[15] For one way of drawing this distinction, see [Smith(2010)].

Φ-assertion. Intuitively, the set will include those alternative contributions that we can expect (in a broadly normative sense) the interlocutor to have considered making instead, at least tacitly or implicitly. For example, $\phi \in \mathcal{ALT}(\phi \wedge \psi)$.[16] It is a notoriously difficult open problem from formal semantics and pragmatics to say exactly what belongs in any given salient alternative set,[17] one that I can't hope to solve here. Nevertheless they are indispensable to basic semantic[18] and pragmatic[19] reasoning. Here are some sufficient conditions for membership in $\mathcal{ALT}(\sigma)$ that will be of use in what follows:

- $\{\phi, \psi\} \subset \mathcal{ALT}(\phi \wedge \psi)$
- $\{\phi, \psi\} \subset \mathcal{ALT}(\phi \vee \psi)$
- $\{\ulcorner$must $\phi\urcorner, \ulcorner$might $\phi\urcorner, \ulcorner$probably $\phi\urcorner\} \subset \mathcal{ALT}(\phi)$
- The null contribution (e.g., silence) $\in \mathcal{ALT}(\sigma)$ for all σ.
- If e is a constituent of σ, and e is on a Horn scale with $e_1, ..., e_n$, then for all σ' such that σ' is like σ except substituting e_1 or ... or e_n for e, $\sigma' \in \mathcal{ALT}(\sigma)$.[20]

This is not an exhaustive list by any means. That being so, it's important to re-emphasize that the salient alternative set will include those sentences that the speaker can be expected easily to have in mind. That is the intuitive criterion I will sometimes appeal to. And if that's really what an alternative set is, rather than something determined by grammar, then we shouldn't expect to be able to specify situation-independent membership criteria for $\mathcal{ALT}$ anyway.

The framework in place, I'll now suggest a necessary condition on rightness, i.e., a sufficient condition on wrongness. Call an utterance of a sentence σ *interest-dominated by* σ' *in* c iff $\sigma' \in \mathcal{ALT}(\sigma)$ and the use of σ in c does at best insignificantly better by every mutual interest in the conversation than does the use of σ' in c, and not merely insignificantly worse by some mutual interest. The use of a sentence σ is *interest-dominated in* c *simpliciter* iff there exists a $\sigma' \in \mathcal{ALT}(\sigma)$ such that the utterance of σ is interest-dominated by σ' in c. Roughly put, the use of a sentence is interest-dominated when there's a salient alternative that the speaker might have said instead that would've done not significantly worse in every way and significantly better in some way.

One aspect of these formulations deserves special comment. Simpler definitions would delete the qualifications involving 'significantly'. Doing that would, however, rob the definitions of any application. If σ is the first sentence our speaker considers and initially elects to utter, there is a

[16] I will be somewhat sloppy with use and mention for readability.

[17] See, e.g., [Katzir(2007)] and [Swanson(2010)].

[18] See, e.g., [Rooth(1992)].

[19] See, e.g., [Abusch(2002)] and [Geurts(2010)], among many others.

[20] A Horn scale is a set of logically related expressions, increasing in strength. For example, {'some', 'most', 'all'} form a Horn scale; $\ulcorner$all Fs are Gs$\urcorner$ entails $\ulcorner$most Fs are Gs$\urcorner$, etc. Other examples are {'one', 'two', ...} and {'warm', 'hot', 'boiling'}. See [Horn(1972)] for the origins of the idea. I make no assumptions about how Horn scales are derived.

small cost to efficiency in switching to some σ'. Similar "costs" are easy to see. Usually, however, they don't matter, since they are dramatically outweighed by most other salient considerations. My official definitions, then, will stand as they are, but sometimes I will speak loosely and drop the qualifications in the interests of efficiency.

Here are two versions of a norm I accept:

NON-DOMINATION NORM (objective). For all S and σ, S must: utter σ in c only if σ is not interest-dominated in c.

NON-DOMINATION NORM (subjective). For all S and σ, S must: utter σ in c only if it's not the case that S should (subjectively) think that σ is interest-dominated in c.

Conversation, remember, is a tool that should used be for advancing our mutual interests. If we speak in a way that only *hinders* those interests compared with others that we easily could've made instead, then we have done something objectively wrong. If we should (subjectively) *think* that the contribution only hinders those interests compared with others that we easily could've made instead, then we have done something subjectively wrong. NON-DOMINATION NORM is the weakest norm that captures that basic idea. That does not mean that other, stronger norms are false. For example, you might have some kind of expected utility norm, according to which speakers should maximize the expected mutual benefit of their contributions. I think this norm is probably too strong, since we are pretty lax about the trade-offs we are willing to accept in conversation. Nevertheless I think that some such stronger norms are also worth consideration. In this section, I just explore what can be done with the weaker NON-DOMINATION NORMS. In later sections I will not assume that the NON-DOMINATION NORMS are the strongest correct norms.

Here are some examples of some interest-dominated sentences that the two norms would rule impermissible in almost any usual situation. (We can always fix special situations where these sentences would be fine things to say, e.g., if someone else said one of them, and you asked me what they said.)

(2) It's raining and it's raining.

(3) It's raining or it's not raining.

(4) I'm either in Paris or France.

(2) will in usual conversations be interest-dominated by:

(5) It's raining.

(5) is in $\mathcal{ALT}((2))$ and communicates the same information that (2) does, but it takes less time and is less strange to hear. (3) will often be interest-dominated by silence, or by many other things a

person might say. Of course, we do sometimes say things like (3), i.e., if we wish to deny indeterminacy. In absence of such purposes, though, (3) is not typically acceptable. Finally, (6) is a *Hurford disjunction*, a disjunction $\ulcorner \phi \vee \psi \urcorner$ such that either ϕ entails ψ or *vice versa*.[21] What *exactly* accounts for their usual infelicity is a matter of open controversy, but personally, the story I like is that (6) usually interest-dominates (4), again by being briefer and less confusing:

(6) I'm in France.

The full explanation of Hurford's constraint is likely to be complicated, both because an algorithm for determining what's in $\mathcal{ALT}((4))$ is likely to be difficult to find, and because other examples seem to show that my initial gloss of Hurford's constraint has to be refined somehow.[22] Here's another example, this time with a little more bite. Suppose that the fact that it's raining is public knowledge among us, indeed in the "active context"—we've been thinking about it recently.[23] Then it is typically inappropriate to say, for most any ϕ:

(7) Suppose it's raining. Then this rain is unusual weather to have in January!

The first part of (7) is a supposition-assertion, using my jargon. But it's useless to get us to *suppose* that it's raining, since we all know full well that it's raining, and that this is manifest to all of us. The supposition is useless at best, but more likely confusing. The speaker should really say something more like "this [rain] is unusual weather to have in January!", and forego the supposition-assertion entirely.

This kind of example leads naturally to my next point. The NON-DOMINATION NORMs do not yet tell us when a Φ-assertion is right or wrong; they only tell us when using a certain kind of sentence is right or wrong. The bridge between them, though, is clear from the discussion of (7): a Φ-assertion is wrong in c if every sentence one could use to make the Φ-assertion is wrong in c. So, a sufficient condition for a given Φ-assertion's being wrong is that there does not exist a sentence σ such that σ is non-dominated in c and using σ in c would constitute a Φ-assertion. What could explain why *every* sentence one could use to impart Φ to one's addressee would be interest-dominated? One reason would be that it's wrong to make their addressee have Φ because of your Φ-assertion, and whatever else the Φ-assertion could accomplish could be accomplished in ways that don't make the addressee have Φ. There might be other cases where the same result—the impermissibility of Φ-asserting—comes about because of a different mechanism. But I'll focus on the former sort of case, since I think it's central.

[21] See [Hurford(1974)].

[22] The problem is that examples like the following seem fine:

(i) He ate some or all of the cookies.

See [Chierchia et al.(2012)Chierchia, Fox, and Spector] for discussion, though I disagree with their account.

[23] For this terminology, see [Kripke(2009)].

The NON-DOMINATION NORMs and the wrongness of making someone have Φ and the availability of alternatives that would do just as well while not making the addressee have Φ entail RIGHTNESS NORM OF Φ-ASSERTION. So, not only does it look like an antecedently attractive view anyway; it also arises very naturally from this framework and very minimal constraints formulated within it.

You might, however, wonder whether we might need *both* FITTINGNESS NORM and RIGHTNESS NORM. After all, as I said, they are compatible. Now, in section 1 I did give some initial counterexamples to the former; but you might wonder whether we can explain the rather large amount of data that seem to support it if we abandon it entirely. In the next section, I'll formulate this challenge in more detail, and show how the rightness theorist can successfully respond to it.

5.3 When We Can Simulate Fittingness Constraints

There are apparently good reasons to think that *even if* something like RIGHTNESS NORM OF Φ-ASSERTION is correct, as I aimed to show in the last section, we'll still need something like FITTINGNESS NORM OF Φ-ASSERTION. Assertion, for example, is epistemically and not merely practically or morally constrained. In this section, I'll try to show two things: first, a rightness theory like the one I presented in the last section can account for epistemic constraints; and two, assertion is far more of a special case than it might have seemed to be.

First, I'll illustrate the challenge. Assertion seems to be epistemically constrained, independent of practical considerations that would otherwise warrant conveying particular pieces of information. As I said in the introduction, it is controversial what the epistemic constraints are. Most famously, some think that someone who asserts that p must know that p, or must express their knowledge in asserting p.[24] But even those who reject any version of the knowledge norm still think that assertion is epistemically constrained. It seems like a condition of adequacy, in fact, that a full theory of the norms of assertion explain how assertions are epistemically constrained. That's because we need to explain the following sorts of data.[25]

- *Lotteries*. If I buy a ticket from an n-person lottery, and I know there will be exactly one winner, then so long as this is *all* I know about the lottery, it is inappropriate for me to say that I have lost the lottery. That can be so even though the probability that I have lost can be arbitrarily close to 1, by making n larger. One explanation is that I cannot know that I have lost, another is that I cannot justifiably believe that I have, and there are still others. What matters, though, is that there be *some* explanation of this.

[24] See, e.g., [Turri(2011)].

[25] I have left out *Moore's paradox*, i.e., that while assertions of the form $\ulcorner\phi$ and I don't [believe/know] $\phi\urcorner$ might be true, they cannot be appropriate. I did so not because I don't think the rightness theorist can explain it, but rather because I think there is primarily a problem with Moorean beliefs rather than assertions. [Sorensen(1988)] is apparently the first person to make this point; see more recently [Coliva(2015)] for interesting discussion.

- *Convertibility*. 'I can't say' and 'I don't know' seem pragmatically equivalent.
- *Prompting and Challenging*. One way for me to ask whether p is to ask whether my interlocutor *knows* whether p. One way for me to challenge my interlocutor's claim that p is to ask whether they know that p.

More generally, you might think, if there are epistemic constraints on belief-assertion, then *any* theory of communicative propriety ought to account for fittingness constraints: that there is something inappropriate about communicating an unfitting mental state. FITTINGNESS NORM of Φ-ASSERTION would, then, be inferred as a generalization from the fact that there are epistemic constraints on assertion. After all, why should belief be special in this regard?

I will show two things: that the rightness theory, using the NON-DOMINATION NORMs, can, under certain conditions, entail the knowledge norm; and second, that belief is *special* in a way that blocks the generalization step in the argument I just sketched.

Suppose my evidence makes it probable or even very, very probable that ϕ (e.g., we're in a lottery case), but where I don't have the right kind of evidence to *warrantedly believe* that ϕ. Here's an important thought of [Williamson(2000)]'s: "Probabilistic evidence warrants only assertion that something is probable" (248). More generally, for any assertion ϕ, our language provides a number of probabilistically hedged alternatives, e.g., ⌜Probably ϕ⌝, or ⌜It's likely that ϕ⌝, or where specific numbers can be attached as in a standard lottery case, ⌜ϕ is $1/n$ likely⌝. These come to mind so easily, and augment an utterance's length so insignificantly, that they will be members of $\mathcal{ALT}(\phi)$. They seem to have no drawbacks relative to simply asserting ϕ, at least on any natural way of filling out the case. They are similar enough to assertions of ϕ that they license much of the same behavior (except perhaps the closing off of inquiry as to whether ϕ, which is a good thing in a case where belief that ϕ would be unwarranted). Will these probabilistic versions also have relative advantages?

They will, I think, at least in ordinary cases. Namely, they will not communicate the *belief that* ϕ. Suppose for a moment that a person who believes that p due to testimony can have warrant for p only if the testifier had warrant for p in testifying that p. So if I communicated that ϕ, I would give my interlocutor an unwarranted belief that ϕ. Now, if an individual *ought not* to believe that ϕ, then it is in their interest—and thus in the interlocutors' mutual interest—not to believe that ϕ. I assume that when S would lack warrant to believe that p, then S ought not to believe that p. There are many possible explanations of why it is that an individual ought not to have unwarranted beliefs, but for my purposes I need not pick between them. In any case, it follows that it is in our interlocutors' mutual interests that the addressee not believe that ϕ. It follows that ϕ is interest-dominated by one of our probabilistic hedges.[26] This is an *epistemic* constraint: the source of the fact that they ought not to believe that ϕ is epistemic. This 'ought'-fact interacts with NON-DOMINATION NORMs and the facts about $\mathcal{ALT}$ that I mentioned in the previous paragraph. *What* epistemic norm we ultimately

[26] Note that ⌜Probably ϕ⌝ and the others can express and communicate warranted beliefs. See, e.g., [Moss(2013)], [Moss(forthcoming)]. (One need not accept her full picture in order to accept this claim.)

get by this procedure depends on what makes a belief unwarranted. If you think that a belief is or would be warranted only if it constitutes or would constitute knowledge,[27] then this process gets you the familiar knowledge norm. From this perspective, the intramural disputes between epistemic theorists about norms of assertion boils down to whether only knowledgeable beliefs are warranted, or other kinds of beliefs as well.

The assumption of the previous paragraph is false, but in a way that in fact helps my overall argument. If, when someone who doesn't know that ϕ, or more generally doesn't believe that ϕ with warrant, but nonetheless in asserting ϕ can make her addressees warranted in believing that ϕ, then it seems not at all wrong for the speaker to assert that ϕ. So, for example, if a creationist high school teacher perfectly conveys the theory of evolution and the evidence that supports it to her students, then it seems like it is perfectly appropriate for her to do so.[28] Rightness theories are, as I said, speaker *and* hearer-focused, unlike fittingness theories. So while fittingness theories struggle with the counterexamples to that assumption, rightness theories do not.

I have, then, explained lotteries—schematically, because I have not said what is required for epistemic warrant. If what I just said was right, then we do, at least in typical cases, have a knowledge norm of assertion (depending on whether warrant requires knowledge). Asking a question will normally make it common knowledge that it is in the mutual interests of the interlocutors that the questioner receive an answer. So, what reason could a speaker have for *not* giving an answer? Well, if they have some implicit understanding of the work of the previous two paragraphs, it will be clear to both interlocutors that the only reason for withholding the answer, in typical cases, will be that the speaker has doubts about whether they know, or warrantedly believe, an answer. Of course, in *some* cases 'I can't say' or 'I can't tell' is *not* interchangeable with 'I don't know'. For example, if you indicate that you've been sworn to secrecy, the former are felicitous but not the latter. But those are simply cases where the speaker makes it clear that the interlocutors' receiving an answer is *not* in their mutual interests. In other words, the rightness theorist is in a *better* position to explain convertibility than a fittingness knowledge norm theorist.

Finally, return to prompts and challenges. Why, first, should I be able to elicit information by asking whether my interlocutor *knows* that ϕ? If I ask that question, the answer must advance my or my interlocutor's interests somehow; but it is unlikely that simply having information about whether you have knowledge whether ϕ is enough—if I cared about that, I would also typically care about the answer. Sometimes we can make it clear we only want to know whether they know whether ϕ, for example when we are doing a survey and it's clear that what matters is whether they think they know whether ϕ. None of this appeals to any specific norms, but in this case, I don't think we need to. Challenges, on the other hand, do. If you say that ϕ, and ϕ seems to me dubious or without sufficient evidential warrant, I can challenge you on that basis, just because, as I argued above, it is typically wrong to communicate that ϕ without sufficient evidential warrant.

[27] See, e.g., [Sutton(2007)], [Bach(2008)], and perhaps [Williamson(2000)].
[28] For these counterexamples to the assumption, see [Lackey(1999)], [Lackey(2007)].

There are other pieces of data in the knowledge norm's favor, but discussing all of them would take me too far afield. For now I think I've supported the following claim: the rightness theorist needn't posit a special fittingness norm to account for these data. Now I want to argue that even if we couldn't pull similar maneuvers with other kinds of Φ-assertions, we don't have to, since the generalization step in the argument was far too quick.

When I illustrated how the rightness theorist could capture something like the knowledge norm, I argued that for any assertion ϕ, there is a nearby hedged version, e.g., $\ulcorner$Probably, $\phi\urcorner$ that interest-dominates ϕ. As I said, $\ulcorner$Probably, $\phi\urcorner$ communicates an attitude very much like the belief that ϕ, but a warranted one. In general, $\mathcal{ALT}(\phi)$ will have sentences in it that interest-dominate ϕ when the belief that ϕ is or would be unwarranted, because the attitudes they communicate are so similar to the belief that ϕ.[29] The important point, however, is that *this is a special feature of belief and assertion*. It is in general not true that, where uttering σ in c would constitute a Φ-assertion, that there are members of $\mathcal{ALT}(\sigma)$ such that uttering one of them in c would constitute the a Ψ-assertion, with Ψ suitably related to Φ. To see the point, return to Bryan's (1):

(1) Where's the beef?

Remember that it is not fitting for Bryan or for his addressees to be amused by (1). But there is in general no $\sigma \in \mathcal{ALT}((1))$ such that uttering σ would communicate amusement. We have no appropriate expectation that someone who makes a bad joke also have considered making a good joke. Good jokes require creativity and effort, not to mention good luck, in a way that probabilistically hedged versions of ϕ do not. What looked like a general constraint on Φ-assertions is actually peculiar to belief, simply for any belief we might like to communicate, there are nearby hedged versions of that belief that are more likely warranted and that we can fairly expect people to have had in mind as alternatives to what they in fact said. Put another way, there are obvious repairs to lottery assertions (and lottery beliefs), but there are no obvious repairs to bad jokes. This is why it looks like assertion obeys a fittingness norm like the knowledge norm and amusement-assertion doesn't. On the present picture, every kind of Φ-assertion obeys exactly the same norms, and the systematic differences in propriety between the different kinds of Φ-assertions arises from how their associated alternative sets work.

This means that even if there are general epistemic constraints on the propriety of assertion, we should not expect there to be general fittingness constraints on Φ-assertion for other sorts of Φ. This blocks the generalization step of the argument with which I began this section. It also leads to a further thought: the propriety of using σ to make a Φ-assertion will often depend on general, structural features of $\mathcal{ALT}(\sigma)$. In particular, where it is easy to improve on fittingness while communicating a Φ-like attitude, then we should find that not improving on fittingness is somehow bad or marked. For some confirmation of this picture, then, it's useful to see whether that prediction is borne out. That's what I'll attempt to do in the rest of this section.

[29] E.g., if Moss and others like [Yalcin(2012)] are right, they communicate the *high credence* that ϕ.

Expressives like ‘damn’ and ‘bastard’ express or communicate an attitude to some content, typically the content the speaker herself expresses. Here are some examples:

(8) My damn dog ate my homework again!

(9) That bastard John stole my car!

(8)’s speaker expresses their negative affect (at least, as of that moment) toward their dog, particularly in response to the dog’s eating their homework. (9)’s speaker expresses negative affect toward John, this affect being presented as at least in part a response to John’s having stolen the speaker’s car. Following [Potts(2007)], expressives have the following properties:

- They are independent from the regular descriptive content of the utterance.
- They concern something of the utterance situation (rather than a past or counterfactual situation).
- They arise from the speaker’s own perspective.
- They are resistant to descriptive paraphrase.
- They are performative (i.e., do their job just by being uttered).
- They intensify with repeated use.

I said that they express or communicate the speaker’s attitude. Sometimes an expressive is used in a way that suggests the speaker’s recognition of the reaction’s idiosyncrasy. But sometimes they are used in order to communicate those attitudes. Suppose I say the following.

(10) That politician is such a fucker.

In saying (10), I aim to communicate my very low opinion of the politician. My speech act was unsuccessful to some extent if you come away with a high opinion of the politician in question. Indeed, if my speech act was totally successful, I can go on to *assume* we take the same attitude toward the politician. I might go on to say something like:

(11) And given that he’s such a fucker, we have to start campaigning against him right now.

One interesting thing about expressives is that they can be graded very finely. ‘Darn’, ‘damn’, and ‘fucking’ form a scale of increasing negative intensity. They form something like a Horn scale, in fact. Other scales might be {‘jerk’, ‘bastard’, ‘fucker’, ...} and {‘cool’, ‘awesome’, ...}. We can even model this formally, if we like.[30] This means that, where e is an expressive in σ, $\mathcal{ALT}(\sigma)$ will often be very rich, just as it was with more normal assertions. That means we can use expressives to

[30] See [Potts(2007)].

test the generalizations that I arrived at in my discussion of belief and assertion, i.e., that fittingness constraints arise from alternative sets with rich enough structure.

So, suppose what prompts me to say (10) is that the politician in question released an attack ad on her opponent. This is, perhaps, a bad thing about the politician, but not *that* bad, in part because it's entirely normal. But even if the politician in question is no doubt a *jerk*, say, because of her attack ad, that doesn't make her bad enough to warrant the epithet in question. If the speaker wanted to communicate her disdain for the politician, she should have used the following:

(12) That politician is such a jerk!

(12) expresses a more fitting attitude given the evidence at the speaker's disposal. The addressee (and the speaker, but here the addressee is the one who matters) ought not to have an unfitting attitude toward the speaker, especially an unfitting negative attitude. Thus it seems to interest-dominate (10). The speaker should have said (12) rather than (10), then, and that is why (10) sounds bad in the situation as described. This is again a fittingness constraint, but one that arises without an explicit fittingness norm, simply because the set of expressives has a rich, gradational structure.

In this section, I've argued that a natural way of saying when a Φ-assertion is wrong, combined with a framework invoking salient alternatives, allows us to capture the data that motivate fittingness theories. This framework makes rightness a function of the interlocutors' interests and the options the speaker should have considered. But that's in general no different from any other sort of action. The framework is natural, then, and allows us to explain a wide range of data while blocking arguments for fittingness theories. We can, in other words, see how *fittingness* constraints, e.g., epistemic constraints, can arise from a general moral framework. We need not posit independent fittingness constraints on Φ-assertions.

5.4 When Fittingness Is No Constraint

I have worked out in some detail how fittingness constraints on Φ-assertions come about. In this section, I will attempt to answer the question with which I ended section 1, in particular to give some sufficient conditions for the absence of fittingness constraints on Φ-assertions. To do that, I'll first introduce a way of classifying different sorts of attitudes.

Bryan's (1) was one example, I think, of when a Φ-assertion was appropriate despite that fact that the amusement it communicated was unfitting. The joy in something totally mundane that one prisoner communicated to the other was another example. In the previous section, I sketched one explanation of why (1) was not inappropriate: $\mathcal{ALT}((1))$ does not have anything that would be fittingly amusing. Let me be a little more explicit. Good jokes are rarely determined by algorithms of the sort that contribute most elements of $\mathcal{ALT}$. So there are likely no *good* jokes in $\mathcal{ALT}((1))$. This isn't *necessarily* the case, mind; if Bryan had nearly told a good joke, but messed up the

punchline somehow, perhaps the good joke really is in $\mathcal{ALT}$. But often those types of cases are simply performance errors, and it's hard to blame or criticize someone for performance errors.

This isn't the whole story, though. It's no accident, I think, that amusement and joy are not only attitudes that might or might not be fitting, but also attitudes that most people have interests in experiencing. Generally speaking, there are attitudes we have "state-given reasons" for having, even if these reasons don't themselves rationalize having those attitudes. Happiness, amusement, and similarly positive emotions work like that, at least much of the time. There are also neutral attitudes, ones that might be instrumentally useful, but for which we have no general state-given reason to have them or not, e.g., belief. Call the first *positive* attitudes, the second *negative* attitudes, and the third *neutral* attitudes.[31] It is, *ceteris paribus*, in interlocutors' mutual interests that they be amused or joyful, against their mutual interests that they be angry or sad, and neither for nor against their mutual interests that they have a given belief. That is not to say that it can't be in someone's interests to have a negative attitude, fittingly or not; it's just to say that intrinsically, something's being a negative attitude is a *pro tanto* reason not to have it.

I conjecture that a Φ-assertion made to S by uttering σ will not have fittingness constraints associated with it if it meets the following conditions:

i. Φ is a kind of positive attitude.

ii. There is no $\sigma' \in \mathcal{ALT}(\sigma)$ such that σ' were one to communicate Φ with σ' to S, S would have Φ more fittingly than S would when Φ is communicated to S with σ, and it would not be wrong to communicate Φ to S with σ'.

I base this conjecture on two things: the goodness of having a positive attitude, and the contrastive nature of communicative propriety, at least on the theory presented here. The thought is, roughly, that unfittingness of an attitude typically has only a minimal effect on one's overall interests, and that an attitude's positivity will *outweigh* its lack of fittingness in light of the fact that there is no salient alternative available to the speaker with which the speaker could fittingly communicate the positive attitude in question. Not everyone will agree with me that unfittingness makes a very small difference to an agent's interests. It is difficult to argue about that sort of claim in general, I think, so I just invite those who disagree to consider Bryan's (1) again: do they *really* think that the unfunniness of the joke compares very much to the value of the amusement, as far as the interlocutors' interests are concerned? For my own part, I don't think they are at all comparable.

The NON-DOMINATION NORMs were not *sufficient* conditions on communicative propriety, but just necessary ones. Nevertheless, I think any plausible way of strengthening them into sufficient conditions (if that could be done) would preserve the goodness of the reasoning in the previous paragraph. For example, an overall utility norm, or an expected utility norm, would validate it,

[31] This distinction, at least in outline, is very old: Aristotle's *Rhetoric* distinguishes between attitudes accompanied by pleasure and attitudes accompanied by pain, for example. The 'state-given' terminology comes from [Parfit(2001)].

as would a range of norms between those and the NON-DOMINATION NORMs. So, though I still consider what I've just argued for a conjecture, it seems to me a plausible one, not just for the reasons I just adduced, but also for its ability to correctly predict the cases from section 1.

Objection. In the cases in section 1, the speakers didn't knowingly communicate an unfitting attitude. Bryan, for example, was amused by the bad joke. But doesn't there seem to be something wrong with intentionally communicating an unfitting attitude, even when the attitude is a positive one? Take the following case, for example:

> *Unfitting happiness*. Jorge does not take decreases in the unemployment rate to be cause for celebration; other indicators, he thinks, are far better guides to how we should feel about the economy. Nevertheless, he thinks Parisa, who *does* take it to be a good indicator of the economic health of the country, could use some cheering up after a rough couple months. So, he makes the following excitement-assertion to her:
>
> (13) How wonderful! The unemployment rate is down this month!

Doesn't it seem as though Jorge has done something impermissible in uttering (13), even though excitement and happiness are positive attitudes? Perhaps even if unfittingness itself isn't enough for communicative impropriety, *knowing* or *intentional* unfittingness is.

Reply. I think there is something wrong with (13), but it isn't quite that Jorge says something knowingly unfitting. (13) is *deceptive*: it makes Jorge sound like he has an attitude that he does not in fact have, namely excitement at the decrease in the country's unemployment rate. It's true that (13) will not be interest-dominated, since excitement is generally a positive attitude. Nevertheless, misrepresenting one's attitudes can be a pretty bad harm, especially if Jorge and Parisa are close and she cares that they be honest with one another. The problem with (13) isn't the knowing lack of fittingness, but the presence of deception, even manipulation. Those can be strong enough harms to very much outweigh whatever benefit Parisa gets from her momentary excitement.

As confirmation of this reply, consider an author who writes "inspirational" biographies of famous people overcoming adversity. If she knows her audience well and cares to cater to them, she might write some schlock that does in fact inspire her audience. This needn't be bad or manipulative: she might make their lives much better, and they don't particularly care one way or the other what she herself thinks about whether the stories really are inspiring. Whether or not these kinds of knowingly unfitting Φ-assertion cases are in fact inappropriate, then, seems to turn on what the actual interests of the individuals involved are, in particular how much they would really care about the speaker's (or author's) misrepresenting themselves. This objection provides support for my account.

5.5 Conclusion

In this chapter, I've tried to develop a rightness-based approach to communicative propriety, one that could simulate fittingness-based approaches when such approaches are correct. My account was contrastive and interest-based, at a very high level of abstraction a form of consequentialism as applied to intentionally causing people to bear given attitudes *via* speech. In doing this, I hope to have shown that we need not posit fittingness norms as theoretical primitives, indeed that we *shouldn't*, since such norms apply only under specific conditions. The account is also much more general than is normally offered, since it applies to every kind of Φ-assertion, not just assertions aimed at communicating beliefs. There remains a great deal of work to be done. To give just one example, I have not said how we can most plausibly strengthen the NON-DOMINATION NORMS, even though they cannot be the full story, as we saw in the last section.

In this area as in so many others, belief has had a large and distorting influence. The norms of expression to which it is subject are idiosyncratic, and so far that fact has inhibited a full understanding of even where epistemic norms of assertion come from. I hope to have at a minimum made a case for thinking that work on communicative propriety requires looking carefully at a wide spectrum of attitudes, even when what's at issue is just belief.

BIBLIOGRAPHY

[Abusch(2002)] Abusch, Dorit. 2002. "Lexical Alternatives as a Source of Pragmatic Presuppositions." In B. Jackson (ed.), *Proceedings of Semantics and Linguistic Theory XII*, 1–19. Ithaca, NY: Cornell.

[Anscombe(2000)] Anscombe, G. E. M. 2000. *Intention*. Cambridge, MA: Harvard. Originally published 1957.

[Arpaly(2002)] Arpaly, Nomy. 2002. *Unprincipled Virtue*. Oxford: Oxford.

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

[Ashwell(2013)] Ashwell, Lauren. 2013. "Deep, Dark...or Transparent? Knowing Our Desires." *Philosophical Studies* 165:245–256.

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

[Bach(2008)] Bach, Kent. 2008. "Applying Pragmatics to Epistemology." *Philosophical Issues* 18:68–88.

[Berker(2013a)] Berker, Selim. 2013a. "Epistemic Teleology and the Separateness of Propositions." *Philosophical Review* 122:337–393.

[Berker(2013b)] —. 2013b. "The Rejection of Epistemic Consequentialism." *Philosophical Issues* 23:363–387.

[Blumberg and Holguín(forthcoming)] Blumberg, Kyle and Holguín, Ben. forthcoming. "Ultra-Liberal Attitude Reports." *Philosophical Studies* .

[Boutilier(2003)] Boutilier, Craig. 2003. "On the Foundations of Expected Expected Utility." In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, 285–290. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

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

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

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

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

[Brogaard(2007)] Brogaard, Berit. 2007. "Sharvy's Theory of Definite Descriptions Revisited." *Pacific Philosophical Quarterly* 88:160–180.

[Broome(1990–1991)] Broome, John. 1990–1991. "Fairness." *Proceedings of the Aristotelian Society* 91:87–101.

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

[Broome(2007)] —. 2007. "Wide or Narrow Scope?" *Mind* 116:359–370.

[Broome(2013)] —. 2013. *Rationality Through Reasoning*. Malden, MA: Wiley-Blackwell.

[Brown(2011)] Brown, Jessica. 2011. "Assertion and Practical Reasoning: Common or Divergent Epistemic Standards?" *Philosophy and Phenomenological Research* 84:123–157.

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

[Burge(1979)] 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.

[Byrne(2011)] Byrne, Alex. 2011. "Knowing What I Want." In JeeLoo Liu and John Perry (eds.), *Consciousness and the Self: New Essays*. Cambridge, UK: Cambridge.

[Cappelen(2011)] Cappelen, Herman. 2011. "Against Assertion." In Jessica Brown and Herman Cappelen (eds.), *Assertion*. Oxford, UK: Oxford.

[Cariani et al.(2013)Cariani, Kaufmann, and Kaufmann] Cariani, Fabrizio, Kaufmann, Magdalena, and Kaufmann, Stefan. 2013. "Deliberative Modality under Epistemic Uncertainty." *Linguistics and Philosophy* 36:225–259.

[Carr(2015)] Carr, Jennifer. 2015. "Subjective *Ought*." *Ergo* 2:678–710.

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

[Chierchia et al.(2012)Chierchia, Fox, and Spector] Chierchia, Gennaro, Fox, Danny, and Spector, Benjamin. 2012. "Scalar Implicature as a Grammatical Phenomenon." In Claudia Maienborn, Klaus von Heusinger, and Paul Portner (eds.), *Semantics: An International Handbook of Natural Language Meaning*, 2297–2331. Berlin: De Gruyter.

[Coliva(2015)] Coliva, Annalisa. 2015. "How to Commit Moore's Paradox." *Journal of Philosophy* 112:169–172.

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

[Dancy(2004)] Dancy, Jonathan. 2004. *Ethics Without Principles*. Oxford, UK: Oxford.

[D'Arms and Jacobson(2000)] D'Arms, Justin and Jacobson, Daniel. 2000. "The Moralistic Fallacy." *Philosophy and Phenomenological Research* 61:65–90.

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

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

[de Sousa(1974)] de Sousa, Ronald B. 1974. "The Good and the True." *Mind* 83:534–551.

[Donnellan(1977)] 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.

[Douven(2006)] Douven, Igor. 2006. "Assertion, Knowledge, and Rational Credibility." *Philosophical Review* 115:449–485.

[Drucker(forthcoming)] Drucker, Daniel. forthcoming. "Policy Externalism." *Philosophy and Phenomenological Research* .

[Egan(2007)] Egan, Andy. 2007. "Epistemic Modals, Relativism, and Assertion." *Philosophical Studies* 133:1–22.

[Engel(2013)] Engel, Pascal. 2013. "Doxastic Correctness." *Proceedings of the Aristotelian Society Supplementary Volume* 87:199–216.

[Evans(1982)] Evans, Gareth. 1982. *The Varieties of Reference*. Oxford: Oxford University Press. Edited by John McDowell.

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

[Feldman and Conee(1985)] Feldman, Richard and Conee, Earl. 1985. "Evidentialism." *Philosophical Studies* 48:15–34.

[Firth(1998)] Firth, Roderick. 1998. *In Defense of Radical Empiricism: Essays and Lectures*. Lanham, MD: Rowman & Littlefield.

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

[Frankfurt(1971)] Frankfurt, Harry G. 1971. "Free Will and the Concept of a Person." *Journal of Philosophy* 68:5–20.

[Frankfurt(1999)] —. 1999. "On Caring." In *Necessity, Volition, and Love*, 155–180. Cambridge, UK: Cambridge.

[Frankfurt(2006)] —. 2006. *Taking Ourselves Seriously and Getting It Right*. Stanford, CA: Stanford.

[Freud(2010)] Freud, Sigmund. 2010. *Civilization and Its Discontents*. New York, NY: Norton. Translated by James Strachey, and originally published 1930.

[Fricker(2009)] Fricker, Miranda. 2009. *Epistemic Injustice*. Oxford, UK: Oxford.

[Friedman(2017)] Friedman, Jane. 2017. "Junk Beliefs and Interest-Driven Epistemology." *Philosophy and Phenomenological Research* 1–16.

[Geurts(2010)] Geurts, Bart. 2010. *Quantity Implicatures*. Cambridge, UK: Cambridge.

[Gibbard(1990)] Gibbard, Allan. 1990. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

[Gibbard(2005)] —. 2005. "Truth and Correct Belief." *Philosophical Issues* 15:338–350.

[Goldman(1987)] Goldman, Alvin I. 1987. "Foundations of Social Epistemics." *Synthese* 73:109–144.

[Graham(2011)] Graham, Peter A. 2011. "'Ought' and Ability." *Philosophical Review* 120:337–382.

[Greaves(2013)] Greaves, Hilary. 2013. "Epistemic Decision Theory." *Mind* 122:915–952.

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

[Grice(1975)] Grice, H. P. 1975. "Logic and Conversation." In Peter Cole and Jerry Morgan (eds.), *Syntax and Semantics, Volume 3: Speech Acts*, 41–58. New York: Academic Press.

[Harman(1986)] Harman, Gilbert. 1986. *Change In View: Principles of Reasoning*. Cambridge, MA: MIT Press.

[Hawthorne et al.(2016)Hawthorne, Rothschild, and Spectre] Hawthorne, John, Rothschild, Daniel, and Spectre, Levi. 2016. "Belief Is Weak." *Philosophical Studies* 173:1393–1404.

[Heck, Jr.(2002)] Heck, Jr., Richard G. 2002. "Do Demonstratives Have Senses?" *Philosophers' Imprint* 2:1–33.

[Henning(2015)] Henning, Tim. 2015. "From Choice to Chance? Saving People, Fairness, and Lotteries." *Philosophical Review* 124:169–206.

[Horn(1972)] Horn, Laurence R. 1972. *On the Semantic Properties of Logical Operators in English*. Ph.D. thesis, UCLA.

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

[Hurford(1974)] Hurford, James. 1974. "Exclusive or Inclusive Disjunction." *Foundations of Language* 11:409–411.

[Hurka(1998)] Hurka, Thomas. 1998. "How Great a Good Is Virtue?" *Journal of Philosophy* 95:181–203.

[Hutcheson(2002)] Hutcheson, Francis. 2002. *An Essay on the Nature and Conduct of the Passions and Affections, with Illustrations on the Moral Sense*. Indianapolis, IN: Liberty Fund.

[Iatridou(1994)] Iatridou, Sabine. 1994. "On the Contribution of Conditional *Then*." *Natural Language Semantics* 2:171–199.

[Jacobson(1995)] Jacobson, Pauline. 1995. "On the Quantificational Force of Free Relatives." In Emmon Bach, Eloise Jelinek, Angelika Kratzer, and Barbara Partee (eds.), *Quantification in Natural Languages*, 451–486. Dordrecht, Netherlands: Kluwer.

[Jeffrey(1965)] Jeffrey, Richard. 1965. *The Logic of Decision*. New York: McGraw-Hill.

[Jeffrey(1986)] —. 1986. "Probabilism and Induction." *Topoi* 5:51–8.

[Jenkins(2015)] Jenkins, C.S.I. 2015. "Modal Monogamy." *Ergo* 2:175–194.

[Johnston(2001)] Johnston, Mark. 2001. "The Authority of Affect." *Philosophy and Phenomenological Research* 61:181–214.

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

[Joyce(1999)] —. 1999. *The Foundations of Causal Decision Theory*. Cambridge, UK: Cambridge.

[Joyce(2009)] —. 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.

[Kagan(1988)] Kagan, Shelly. 1988. "The Additive Fallacy." *Ethics* 99:5–31.

[Kamtekar(2006)] Kamtekar, Rachana. 2006. "Plato on the Attribution of Conative Attitudes." *Archiv für Geschichte der Philosophie* 88:127–162.

[Kaplan(1968)] Kaplan, David. 1968. "Quantifying In." *Synthese* 19:178–214.

[Kaplan(1978)] —. 1978. "Dthat." In Peter Cole (ed.), *Syntax and Semantics*, 221–243. Academic Press.

[Kaplan(1989)] —. 1989. "Demonstratives." In Joseph Almog, John Perry, and Howard Wettstein (eds.), *Themes from Kaplan*, 481–563. New York: Oxford University Press.

[Katzir(2007)] Katzir, Roni. 2007. "Structurally-Defined Alternatives." *Linguistics and Philosophy* 30:669–690.

[Kavka(1983)] Kavka, Gregory S. 1983. "The Toxin Puzzle." *Analysis* 43:33–36.

[Kolodny(2005)] Kolodny, Niko. 2005. "Why Be Rational?" *Mind* 114:509–563.

[Kolodny(2007)] —. 2007. "How Does Coherence Matter?" *Proceedings of the Aristotelian Society* 107:229–263.

[Kolodny and MacFarlane(2010)] Kolodny, Niko and MacFarlane, John. 2010. "Ifs and Oughts." *Journal of Philosophy* 107:115–143.

[Korsgaard(2008)] Korsgaard, Christine M. 2008. "The Myth of Egoism." In *The Constitution of Agency*, 69–99. Oxford, UK: Oxford.

[Kripke(1972)] 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).

[Kripke(1979)] —. 1979. "A Puzzle About Belief." In Avishai Margalit (ed.), *Meaning and Use*, 239–83. Dordrecht: Reidel.

[Kripke(2009)] —. 2009. "Presupposition and Anaphora: Remarks on the Formulation of the Projection Problem." *Linguistic Inquiry* 40:367–386.

[Lackey(1999)] Lackey, Jennifer. 1999. "Testimonial Knowledge and Transmission." *Philosophical Quarterly* 49:471–90.

[Lackey(2007)] —. 2007. "Norms of Assertion." *Noûs* 41:594–626.

[Langton(2012)] Langton, Rae. 2012. "Beyond Belief: Pragmatics in Hate Speech and Pornography." In Ishani Maitra and Mary Kate McGowan (eds.), *Speech and Harm*, 72–93. Oxford, UK: Oxford.

[Lauria(2014)] Lauria, Federico. 2014. *The Logic of the Liver: A Deontic View of the Intentionality of Desire*. Ph.D. thesis, University of Geneva.

[Levinson(2003)] Levinson, Dmitry. 2003. "A Probabalistic Model-Theoretic Semantics for *Want*." In R. Young and Y. Zhou (eds.), *SALT XIII*, 222–239.

[Lewis(1973)] Lewis, David. 1973. *Counterfactuals*. Oxford: Blackwell.

[Lewis(1981)] —. 1981. "What Puzzling Pierre Does Not Believe." *Australasian Journal of Philosophy* 59:283–289.

[Lewis(1988)] —. 1988. "Desire as Belief." *Mind* 97:323–332.

[Link(1983)] 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(2014)] MacFarlane, John. 2014. *Assessment Sensitivity*. Oxford: Oxford.

[Maitra(2011)] Maitra, Ishani. 2011. "Assertion, Norms, and Games." In Jessica Brown and Herman Cappelen (eds.), *Assertion*, 277–296. Oxford, UK: Oxford.

[Maitra and Weatherson(2010)] Maitra, Ishani and Weatherson, Brian. 2010. "Assertion, Knowledge, and Action." *Philosophical Studies* 149:99–118.

[McGee(1985)] McGee, Vann. 1985. "A Counterexample to *Modus Ponens*." *Journal of Philosophy* 82:462–471.

[McKinsey(2009)] McKinsey, Michael. 2009. "Thought by Description." *Philosophy and Phenomenological Research* 78:83–102.

[Mehta(2016)] Mehta, Neil. 2016. "Knowledge and Other Norms for Assertion, Action, and Belief: A Teleological Account." *Philosophy and Phenomenological Research* 93:681–705.

[Miller(2014)] Miller, Kristie. 2014. "Conditional and Prospective Apologies." *Journal of Value Inquiry* 48:403–417.

[Montague(2007)] Montague, Michelle. 2007. "Against Propositionalism." *Noûs* 41:503–518.

[Moran(2001)] Moran, Richard. 2001. *Authority and Estrangement*. Princeton, NJ: Princeton.

[Moss(2012)] Moss, Sarah. 2012. "Updating as Communication." *Philosophy and Phenomenological Research* 85:225–248.

[Moss(2013)] —. 2013. "Epistemology Formalized." *Philosophical Review* 122:1–43.

[Moss(forthcoming)] —. forthcoming. *Probabilistic Knowledge*. Oxford, UK: Oxford.

[Nussbaum(1979)] Nussbaum, Martha. 1979. "The Speech of Alcibiades: A Reading of Plato's *Symposium*." *Philosophy and Literature* 3:131–172.

[Parfit(1984)] Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

[Parfit(2001)] —. 2001. "Rationality and Reasons." In Daniel Egonsson, Jonas Josefsson, Björn Petersson, and Toni Rønnow-Rasmussen (eds.), *Exploring Practical Philosophy: From Action to Values*, 17–39. Aldershot, UK: Ashgate.

[Paul(2014)] Paul, L. A. 2014. *Transformative Experiences*. Oxford, UK: Oxford.

[Penner(1991)] Penner, Terry. 1991. "Desire and Power in Socrates: The Argument of *Gorgias* 466A-468E that Orators and Tyrants Have No Power in the City." *Apeiron* 24:147–202.

[Pettigrew(2016)] Pettigrew, Richard. 2016. *Accuracy and the Laws of Credence*. Oxford, UK: Oxford.

[Pettit and Smith(1996)] Pettit, Philip and Smith, Michael. 1996. "Freedom in Belief and Desire." *Journal of Philosophy* 93:429–449.

[Plantinga(1976)] Plantinga, Alvin. 1976. *God, Freedom, and Evil*. Grand Rapids, MI: Eerdmans.

[Pollock and Cruz(1999)] Pollock, John and Cruz, Joseph. 1999. *Contemporary Theories of Knowledge*. Towota, NJ: Rowman & Littlefield.

[Potts(2007)] Potts, Christopher. 2007. "The Expressive Dimension." *Theoretical Linguistics* 33:165–198.

[Putnam(1975)] 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(1956)] Quine, W. V. O. 1956. "Quantifiers and Propositional Attitudes." *Journal of Philosophy* 53:177–87.

[Quinn(1992)] Quinn, Warren. 1992. "Rationality and the Human Good." *Social Philosophy and Policy* 9:81–95.

[Rabinowicz and Rønnow-Rasmussen(2004)] Rabinowicz, Wlodek and Rønnow-Rasmussen, Toni. 2004. "The Strike of the Demon." *Ethics* 114:391–423.

[Railton(1984)] Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy & Public Affairs* 13:134–171.

[Regan(1980)] Regan, Donald. 1980. *Utilitarianism and Cooperation*. Oxford: Oxford.

[Rescher(1959)] Rescher, Nicholas. 1959. "Choice without Preference." *Kant-Studien* 51:142–175.

[Rooth(1992)] Rooth, Mats. 1992. "A Theory of Focus Interpretation." *Natural Language Semantics* 1:75–116.

[Rosenberg(1984)] Rosenberg, David. 1984. "The Causal Connection in Mass Exposure Cases: A "Public Law" Vision of the Tort System." *Harvard Law Review* 97.

[Ross and Schroeder(2014)] Ross, Jacob and Schroeder, Mark. 2014. "Belief, Credence, and Pragmatic Encroachment." *Philosophy and Phenomenological Research* 88:259–288.

[Russell(1912)] Russell, Bertrand. 1912. *The Problems of Philosophy*. London: Williams and Norgate. Paperback edition by Oxford University Press, 1959.

[Schelling(1960)] Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard.

[Schiffer(1977)] Schiffer, Stephen. 1977. "Naming and Knowing." *Midwest Studies in Philosophy* 2:28–41.

[Schoenfield(2015)] Schoenfield, Miriam. 2015. "Bridging Rationality and Accuracy." *Journal of Philosophy* 112:633–657.

[Sharvy(1980)] Sharvy, Richard. 1980. "A More General Theory of Definite Descriptions." *Philosophical Review* 89:607–624.

[Silk(2014)] 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.

[Singer(2009)] Singer, Peter. 2009. *The Life You Can Save*. New York, NY: Random House.

[Smith(2010)] Smith, Holly M. 2010. "Subjective Rightness." *Social Philosophy and Policy* 27:64–110.

[Smith(1994)] Smith, Michael. 1994. *The Moral Problem*. Oxford: Basil Blackwell.

[Smith(2004)] —. 2004. "In Defence of *The Moral Problem*: A Reply to Brink, Copp, and Sayre-McCord." In *Ethics and the* A Priori, 259–296. Cambridge, UK: Cambridge.

[Sorensen(1988)] Sorensen, Roy. 1988. *Blindspots*. Oxford: Oxford University Press.

[Srinivasan(forthcoming)] Srinivasan, Amia. forthcoming. "The Aptness of Anger." *Journal of Political Philosophy* .

[Stalnaker(1978)] Stalnaker, Robert. 1978. "Assertion." In Peter Cole and Jerry Morgan (eds.), *Syntax and Semantics, Volume 9: Pragmatics*, 315–32. New York: Academic Press.

[Stalnaker(1968)] Stalnaker, Robert C. 1968. "A Theory of Conditionals." In *Studies in Logical Theory: American Philosophical Quarterly Monograph Series, No. 2*. Oxford: Blackwell.

[Stalnaker(2008)] —. 2008. *Our Knowledge of the Internal World*. Oxford, UK: Oxford.

[Stampe(1987)] Stampe, Dennis W. 1987. "The Authority of Desire." *Philosophical Review* 96:335–381.

[Steinbock(1986)] Steinbock, Bonnie. 1986. "Adultery." *QQ: Report from the Center for Philosophy and Public Policy* 6:12–14.

[Sutton(2007)] Sutton, Jonathan. 2007. *Without Justification*. Cambridge, MA: MIT.

[Swanson(2010)] Swanson, Eric. 2010. "Structurally Defined Alternatives and Lexicalizations of XOR." *Linguistics and Philosophy* 33:31–36.

[Swanson(2016)] —. 2016. "The Application of Constraint Semantics to the Language of Subjective Uncertainty." *Journal of Philosophical Logic* 45:121–146.

[Turri(2011)] Turri, John. 2011. "The Express Knowledge Account of Assertion." *Australasian Journal of Philosophy* 89:37–45.

[Turri(2016)] —. 2016. *Knowledge and the Norm of Assertion*. Cambridge, UK: Open Book.

[Ullmann-Margalit and Morgenbesser(1977)] Ullmann-Margalit, Edna and Morgenbesser, Sidney. 1977. "Picking and Choosing." *Social Research* 44:757–788.

[Unger(1975)] Unger, Peter. 1975. *Ignorance*. Oxford: Oxford University Press.

[van Fraassen(1980)] van Fraassen, Bas. 1980. "Review of Brian Ellis, *Rational Belief Systems*." *Canadian Journal of Philosophy* 10:193–197.

[Vendler(1972)] Vendler, Zeno. 1972. *Res Cogitans*. Ithaca, NY: Cornell.

[Villalta(2008)] Villalta, Elisabeth. 2008. "Mood and Gradability: An Investigation of the Subjunctive Mood in Spanish." *Linguistics and Philosophy* 31:467–522.

[Vlastos(1973)] Vlastos, Gregory. 1973. "The Individual as an Object of Love in Plato." In *Platonic Studies*, 3–42. Princeton, NJ: Princeton.

[von Fintel(2012)] 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(2002)] Wedgwood, Ralph. 2002. "Internalism Explained." *Philosophy and Phenomenological Research* 65:349–369.

[Wedgwood(2014)] —. 2014. "Objective and Subjective 'Ought'." Manuscript.

[Weiner(2005)] Weiner, Matthew. 2005. "Must We Know What We Say?" *Philosophical Review* 114:227–251.

[Weirich(1980)] Weirich, Paul. 1980. "Conditional Utility and Its Place in Decision Theory." *Journal of Philosophy* 77:702–715.

[Williams(1981a)] Williams, Bernard. 1981a. "Internal and External Reasons." In *Moral Luck*, 101–113. Cambridge: Cambridge.

[Williams(1981b)] —. 1981b. "Persons, Character, and Morality." In *Moral Luck*, 1–19. Cambridge, UK: Cambridge.

[Williamson(2000)] Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

[Worsnip(forthcoming)] Worsnip, Alex. forthcoming. "The Conflict of Evidence and Coherence." *Philosophy and Phenomenological Research* .

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

[Zimmerman(2011)] Zimmerman, Michael J. 2011. "Partiality and Intrinsic Value." *Mind* 120:447–483.