Doxastic Normativity

by

Daniel J. Singer

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2012

Doctoral Committee:

Professor James M. Joyce, Chair
Professor Allan Gibbard
Assistant Professor Sarah Moss
Assistant Professor Chandra Sekhar Sripada

ACKNOWLEDGEMENTS

I've learned a lot very quickly in graduate school. Final versions of philosophy articles often appear as though they were created by a meticulous genius. The genius starts by carefully crafting each argument and then lays them out one by one, in a perfectly reasoned and anticipated fashion. In graduate school, it has become clear that, except for maybe David Lewis, papers are rarely produced like that. Instead, we often start out with a kernel of an idea and, by scraping it on the wits of those around us, slowly build it into something worth considering. I am extremely grateful to the faculty and students of Philosophy at the University of Michigan, who have help me learn how to more closely approximate the meticulous genius, particularly including my dissertation committee, James M. Joyce (chair), Allan Gibbard, Sarah Moss, and Chandra Sripada. I am also extremely grateful to many others outside of the Department of Philosophy at Michigan who have contributed to my philosophical training so far, including in particular Michael Weisberg, Patrick Grim, Daniel Nolan, and the Departments of Philosophy and visitors at the Australian National University and the University of Pennsylvania, where I have spent some time.

The chapters presented below were developed mostly independently of each other, so different collections of people were influential in producing each one. The core idea presented in Chapter I was developed in conversations with Alex Silk. I am extremely grateful to him for his help on this. I also want to thank Allan Gibbard, James M. Joyce, Sarah Moss, Peter Railton, Chandra Sripada, W. Robert Thomas, and the University of Michigan Graduate Student Working Group for feedback on previous versions of that work. For Chapter II, I'm thankful to a whole bunch of awesome people for their help on developing the ideas presented, particularly including Alex Silk, J. Dmitri Gallow, the University of Michigan Graduate Student Working Group, James M. Joyce, Sarah Moss, Chandra Sripada, and Allan Gibbard. Chapter III has seen many versions and forms of presentation. I'm very thankful to Sarah Buss, David Chalmers, Nathaniel Coleman, J. Dmitri Gallow, Allan Gibbard, Patrick Grim, Warren Herold, James M. Joyce, Jason Konek, Maria Lasonen-Aarnio,

Sarah Moss, Stephen Nayak-Young, Daniel Nolan, Daniel Peterson, Peter Railton, Tad Schmaltz, Jonathan Shaheen, Patrick Shirreff, Alex Silk, Chandra Sripada, Aarti Surti, Michael Weisberg, and audiences at the University of Michigan, the University of Southern California, the University of Pennsylvania, Yale University, and the Australian National University for their help in developing my thoughts on those topics. For their help in formulating the ideas presented in Chapter IV, I am very thankful to many people, including James M. Joyce, Jason Konek, Alex Silk, Allan Gibbard, Sarah Moss, Chandra Sripada, David Plunkett, and the University of Michigan Formal Epistemology Working Group. I am particularly thankful to James M. Joyce and Allan Gibbard for introducing me to much of the background to these topics and extensively helping me to reconstruct the ideas in that literature.

I am also very thankful for the financial and research support, which has facilitated my completion of this dissertation, provided in primary part by the University of Michigan, including the Department of Philosophy and the Rackham Graduate School.

Finally, but probably most importantly, I'm very thankful to my family and friends who have supported me through this process, particularly including my mom Lori Singer, her sisters Abbey and Fran Kessman, W. Robert Thomas, Alex Silk, Dmitri Gallow, Ruth McAdams, and Aarti Surti. I also want to thank all of the people who have been supportive to me but who I've forgotten to explicitly mention here.

TABLE OF CONTENTS

CHAPTER

ABSTRACT

Doxastic Normativity
by
Daniel J. Singer

Chair: James M. Joyce

Some 'ought' claims are practical: If you want to make guacamole for dinner, you ought to buy some avocados. Other 'ought' claims are moral: you ought not steal the avocados. Still other 'ought' claims are epistemic: when you see a sign that says the avocados are sold out, you ought to believe there aren't any avocados for sale. This final *epistemic* 'ought' claim doesn't seem to be moral or dependent on your desires, like the other two are. This dissertation is an investigation into the nature of these kinds of 'ought' claims, the oughts for belief.

The first two chapters of this dissertation focus on a puzzle about correct belief. The puzzle is about Hume's dictum that no 'ought' follows from an 'is' – the putative *is-ought gap* or *autonomy of ethics*. From the premise 'Snow is white,' we can infer 'Sophia's belief that snow is white is correct.' But, 'Snow is white' is paradigmatically non-normative, and that Sophia's belief is correct, which is a claim about what belief she ought to have, seems to be normative. Moreover, the argument seems valid, so the is-ought gap is supposed to block this kind of inference. The puzzle is over whether we should give up the autonomy of ethics or find another way to resolve the apparent conflict.

To even make sense of the puzzle, we must have a clear understanding of how the is-ought gap is supposed to work. A.N. Prior showed in 1960 that a simple formulation of the is-ought gap obviously fails. In the first chapter, I consider some ways of reformulating it. I suggest that focusing on the syntax of arguments for a solution, as some have tried, is misguided. Instead, we should look for a semantics-based approach to understanding the autonomy claim. One such approach is from Russell and Restall and employs models of deontic logic. I provide a new

formulation of the autonomy of ethics in a more general semantic framework. A downside of my approach is that it doesn't treat the autonomy claim as blocking inferences from one class of sentences to another. But, my approach is more general than Russell and Restall's, and it offers an explanation of what goes wrong in purported counterexamples to the autonomy claim that the other account lacks.

With that understanding of the is-ought gap in hand, I turn in chapter 2 to discussing the puzzle about correct belief. Allan Gibbard claims that by understanding how objective and subjective oughts relate to each other, we'll see that the offending argument doesn't really bridge the is-ought gap. I cast doubt on Gibbard's proposed solution by claiming that no epistemology of the normative fits nicely with it. Any epistemology that Gibbard could supply would either be unable to explain an awkward asymmetry it must posit in the subjective ought or make subjective oughts too detached from the agents they apply to. In arguing for this disjunction, I also argue that the subjective ought does not play the role in our deliberation that Gibbard expects it to. Really, there is no unique ought that factors in our deliberation. We should reject Gibbard's proposed solution to the puzzle, I claim, because it gets wrong the relationship between objective and subjective oughts. I suggest another solution to the puzzle, one that takes the concept BELIEF to be normative.

In chapter 3, I turn to defend the solution suggested in the previous chapter while more directly tackling the question of the nature of oughts for belief. I offer a new explanation of why we ought to believe the truth. At the heart of the account is the idea that the concept BELIEF is normative. It's a conceptual truth, I claim, that other things being equal, beliefs ought to be true. In support of this, I show how it can give a compelling explanation of an aspect of knowledge ascriptions that is otherwise difficult to explain. I then claim that being an agent requires being subject to this norm of belief. This results in a non-moral, distinctly doxastic, account of why we ought to believe the truth. My conclusion is that asking why we ought to believe the truth is like asking why a bachelor must be unmarried: the answer is contained in the ideas that make up the question.

In the final chapter, I respond to Gibbard's claim that an analogous story cannot work for 'ought' claims for degreed belief. In "Rational Credence and the Value of Truth," Gibbard makes two key claims: (1) that epistemic rationality cannot be understood in terms of the aim of belief, and (2) that epistemic rationality is more plausibly understood in terms of a practical notion, that of maximizing prospective guidance value. When we clarify the role of idealizations in Gibbard's explanation,

we see that Gibbard is left appealing to merely hypothetical bets to explain epistemic rationality even in ideal cases. I claim that these merely hypothetical bets don't seem to be able to play the role he needs them to though and Gibbard must explain how these explanations work. I then undermine Gibbard's motivation for pursuing such an account in the first place, by showing that the argument for the first claim is unsound. Gibbard isn't using the best conception of the aim of belief. I survey five senses in which something or someone can be said to 'aim.' According to the normative sense in which 'belief aims,' belief aims at the truth because having a true content is the standard of good or correct belief. By taking belief to have an aim in this way, I sketch how we can make sense of epistemic rationality in terms of that aim.

CHAPTER I

Mind the Is-Ought Gap

ABSTRACT

Hume supposedly taught us that 'ought's don't follow from 'is's – that's the Autonomy of Ethics. It has been known for a while that a simple way to make sense of Hume's claim fails. In this chapter, I consider some ways of responding to the worry. I suggest that the way that some have tried to save the Autonomy of Ethics, in terms of the syntax of arguments, is misguided. Instead, we should look for a semantics-based approach to understanding the autonomy claim. One such approach is from Russell and Restall and employs models of deontic logic. I provide a new formulation of the autonomy of ethics in a more general semantic framework. A downside of my approach is that it doesn't treat the autonomy claim as blocking inferences from one class of sentences to another. But, my approach is more general than Russell and Restall's, and it offers an explanation of what goes wrong in purported counterexamples to the autonomy claim that the other account lacks.

A lesson we learned from Hume is often summed up as "no ought from an is." Roughly, the idea is that how things are doesn't determine how things ought to be. This is the putative *autonomy of ethics* or *is-ought gap*.

The autonomy of ethics was challenged by several compelling hard cases from Prior (1960a). Then defenses and modifications of the doctrine were proposed, including those by Jackson (1974) and Pigden (1989).

In this chapter, I'll motivate a new way to think about the autonomy of ethics. The is-ought gap is best understood in terms of the semantics, rather than the syntax, of arguments, I'll claim. Russell and Restall (2010) give one such account of the autonomy claim in semantic terms. I'll provide a new account in more general semantic terms that is much simpler than their account and has some advantages over their approach.

1.1 The Autonomies of Ethics

In book III, part I, section I of the *Treatise of Human Nature*, Hume gives the first statement of the autonomy of ethics. He says:

> In every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary ways of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when all of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not. This change is imperceptible; but is however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given; for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it. (Hume 1978, p. 468)

Inferring normative facts from only descriptive facts, Hume seems to be claiming, is unwarranted because "this new relation" of normativity seems "entirely different" from descriptive or observational propositions. What *is* the case, the observational or descriptive facts, do not seem to impinge at all on how things *ought* to be. This simple doctrine is a first rough pass at the autonomy of ethics:

SIMPLE AUTONOMY The normative isn't determined by anything non-normative.

SIMPLE AUTONOMY, if right, would have it that no way of fixing the descriptive facts is enough to figure out how things ought to be. It follows from this that we can't properly argue from how things are to how they ought to be – that is, there is no valid argument with only descriptive premises and a non-trivial normative conclusion.

1.1.1 Two Conceptions of the Simple View

When Hume introduces the is-ought gap, he points to the arguments of his contemporaries that make moral conclusions from descriptive premises. He worries about whether there really is that link from the descriptive to the normative that those arguments need. The normative part of the arguments "expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given; for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it," he (1978, p. 468) tells us.

It seems to me that Hume's worry, as stated, is ambiguous between an *epistemic* reading and a *metaphysical* reading. On the epistemic reading, the claim is about what we can properly *infer*: We cannot draw any conclusions about what ought to be the case just from what is the case, because such an argument would be semantically invalid or because it would be bad reasoning in another way. The metaphysical reading expresses a deeper disconnection between the normative and the non-normative. On the metaphysical reading, what is the case normatively isn't fixed by non-normative facts. If there were a creator who set up all of the facts, the metaphysical version of the autonomy of ethics would demand that the creator set the normative facts even after all of the non-normative facts were established.

Strictly speaking, both versions of the claim also admit of degreed readings: The extreme version of the epistemic claim says that we can reach *no* normative conclusions from non-normative premises. Less extreme versions claim that there are at least *some* normative questions epistemically left open by the non-normative facts. The analogous degreed understanding applies to the metaphysical version of the autonomy claim: According to most extreme version, no normative facts are fixed by the non-normative facts. On less extreme metaphysical readings of the claim, there are least some normative questions metaphysically left open by the non-normative facts.

There are also natural connections between the two versions of the autonomy claim. For example, if the extreme version of the metaphysical claim is correct, it would imply that one cannot properly argue from just non-normative facts to any normative facts (without any bridge principles). On the other hand, if it's possible to reason from some non-normative sentences to some normative ones, then the metaphysical claim would fail as well.

Since it has received the most attention in the literature, in this chapter, I'll focus exclusively on the most extreme versions of the epistemic claim, the claim that *we cannot properly reason* from claims about how things are to a claim about how things ought to be.[1]

[1] For more on the metaphysical version of the claim, see Blackburn's (1971) discussion of the

1.1.2 Problems with the Simple View

A. N. Prior (1960a) shows that the epistemic reading of SIMPLE AUTONOMY is false with three examples:[2]

(A) 1. Tea-drinking is common in England.
2. Therefore, either tea-drinking is common in England or all New Zealanders ought to be shot.

(B) 1. There is no man over 20ft high.
2. Therefore, there is no man over 20ft high who ought to sit in an ordinary chair.

(C) 1. Undertakers are church officers.
2. Therefore, if church officers ought to be reverent, undertakers ought to be reverent.

Each of these examples is a logically valid argument with non-normative premises and a conclusion that contains a normative term essentially.[3] So, the examples show that there are formally valid arguments from non-normative claims to normative conclusions that are valid solely in virtue of the semantics of the connectives. In doing so, they show the failure of the epistemic reading of SIMPLE AUTONOMY: we can properly reason from non-normative premises to a normative conclusion.[4]

There seems to be something deficient about each of the conclusions of these argument. One might think that we could easily avoid Prior's worries by identifying that deficiency and omitting similarly deficient conclusions from the intended scope of the autonomy of ethics.

Prior considers this move, but he argues that it won't work by showing that each of the conclusions can be used in the context of an argument where it appears to have interesting, possibly action-guiding, normative force. For example, we can deduce the conclusion of the first argument like this:

1. Anyone who does what is not common in England ought to be shot;

supervenience of the normative on the non-normative and Jackson's (2013) remarks on how this affects the metaphysical claim.

[2] Here I actually provide the slightly modified versions from Pigden (1989, p. 132) of each of Prior's examples.

[3] A term is essential when it cannot in general be replaced with another term of the same grammatical category without a change in truth-value.

[4] Though Prior did not think of a claim like SIMPLE AUTONOMY as ambiguous between the two given readings, they also show against the metaphysical reading of the claim: since there are valid arguments from non-normative facts to some normative facts, that means that some non-normative facts must fix some normative facts.

2. All New Zealanders drink tea;
3. Therefore either tea-drinking is common in England or all New Zealanders ought to be shot.

This example shows that, though we may have the feeling that there's no real force behind the normative terms in the conclusion of the first argument, it is not normatively impotent. Prior concludes then that there is no is-ought gap.

1.2 Avoiding Counterexamples to the Epistemic Version?

Several responses to Prior's worries for the epistemic version of the autonomy claim have been given in the literature. One popular response is to limit the intended domain of the autonomy of ethics to sufficiently simple normative conclusions. The underlying idea is that none of Prior's examples are compelling counterexamples to the is-ought gap because none of them entail a simple statement of what ought or ought not be the case. Gibbard (2012, p. 80-81), for example, argues that it's quite commonplace for hypothetical ought claims to follow from descriptive premises, so the the autonomy claim should be restricted to sufficiently simple conclusions. Gibbard (2012, p. 88) later contends that the specific puzzle he is considering does not pose a problem for the autonomy of ethics when restricted to simple normative conclusions, where a normative conclusion is *simple* when its main connective is the primitive subjective ought that Gibbard is interested in.[5]

The kind of move that Gibbard employs, which restricts the conclusions that the autonomy claims is meant to apply to, is both too strong and too weak. First, the move is too strong because it excludes some arguments from the purview of the autonomy of ethics that it should not. For instance, it excludes an argument with the conclusion "Either Jane ought to eat tomato soup, or Ange ought to buy garlic bread." This conclusion doesn't entail any simple normative sentence in Gibbard's sense, but surely the autonomy claim is meant to rule out the validity of arguments with only non-normative premises and this conclusion.

Note that to show that Gibbard's restriction is too strong, I need not actually produce a valid argument from descriptive premises to this conclusion (as this would also pronounce on the fate of the autonomy of ethics, not merely clarify its intended domain). Rather, to show Gibbard's restriction to be too strong, I must

[5] Note that Gibbard does take himself to be responding to Prior here. Rather, he is considering some potential outs for the puzzle concerning correct belief discussed below. He does conclude though that "We can maintain that no elementary ought follows analytically from an is" Gibbard (2012, p. 88) in light of his consideration of that puzzle.

only show that if there were such an argument, it would be a counterexample to the autonomy of ethics. But surely an argument with only non-normative premises and the given disjunctive conclusion "expresses some new relation or affirmation" and "'tis necessary that it shou'd be observ'd and explain'd."

Thinking of the autonomy of ethics as only applying to arguments with simple normative conclusions also seems to be too weak. Russell (2010, p. 5) attributes the following example to Gideon Rosen: Suppose "to flurg" means to do something that one ought not do in front of children. Then we could deduce from "Lauren is in front of children." to "Lauren ought not flurg." The conclusion here is simple, in Gibbard's sense, but it also follows from the non-normative premise.[6] So, modifying the domain of the autonomy of ethics to make it apply only to simple normative sentences is too strong in that it doesn't capture all of the intended domain of the autonomy of ethics, and it is too weak in that it doesn't save the claim of Prior-style counterexamples like this one from Russell (2010, p. 5). Other modifications must be sought.

In Prior's original statement of his examples, he notices that the normative terms in his first two examples seem to appear to do no particularly normative work; in them, "the duty established is not one that we need ever be practically anxious about" (1960a, p. 203). Prior diagnoses this as what he calls "contingent vacuousness". A term is contingently vacuous when it can be replaced with a grammatical counterpart without sacrificing the validity of the argument. Prior's first two examples would be still be valid if 'ought' where replaced by 'want', for example. Prior fears that this cannot be the feature that sets his counterexamples to the is-ought gap apart from other arguments though, because he thinks that the 'ought' in his third example is not contingently vacuous (1960a, p. 204-206).

Jackson (1974) considers examples like Prior's first two and diagnoses them as having a property similar to Prior's contingent vacuousness, that of being *ethically invariant*. An argument is ethically invariant when the normative terms in it can be replaced uniformly with any other grammatically appropriate term without sacrificing argument validity (1974, p. 91).[7] When an argument is ethically invariant,

[6] One might worry that stipulating concepts like FLURG is illegitimate and subject to worries like those surrounding the concept TONK as discussed by Prior (1960b). In general, that seems right, but this particular case is not worrisome. FLURG is plausibly quite similar to everyday concepts such as BEING OBNOXIOUS: being obnoxious requires satisfying some descriptive criteria, such as being loud or unrelenting, as well as being such that one ought not be that way. The example of 'flurg' is used since it's easier to specify in the stipulated case what exactly the descriptive content is.

[7] This is a slightly different property than Prior's contingent vacuousness; see Pigden (1989, p. 133-134) for an explanation of the difference and discussion of it.

its validity is, in a sense, independent of the ethical terms. Jackson considers this to be a sufficient condition for being outside of the scope of the autonomy of ethics. If the argument would be valid no matter which term is there, why would having a normative term there be threatening in the way that "'tis necessary that it shou'd be observ'd and explain'd"?

Following this line, Jackson proposes a new conception of the autonomy of ethics that's meant to avoid Prior's examples:

JACKSON AOE No argument from factual premises to an ethical conclusion is valid unless it is either factual-invariant or ethical-invariant (or both), or is reducible to such by synonymy substitution. (1974, p. 93)

where an argument is Σ-invariant when its validity is independent of the meaning of the terms of type Σ (1974, p. 91).

Pigden (1989, p. 134-135), following MacIntyre (1981, p. 54-55), objects to Jackson's method:

> If the redefined autonomy of ethics is not derived from some general logical principle, what reason do we have to believe it true, besides the, perhaps temporary, drying up of counterexamples? In Lakatosian terms Jackson (and in effect Shorter) construct an exception-barring definition which marks out a (hopefully) safe domain for the modified conjecture.

With this motivation, Pigden sets out to provide an argument for the safety of the is-ought gap in (a class of arguments strictly wider than)[8] Jackson's class of arguments. Pigden argues that it can be shown on the basis of the conservativeness of logic that a valid argument from non-normative premises can only contain 'ought' (conceived of as a predicate) in its conclusion vacuously. I will skip most of the details of that argument here,[9] but notice that the possibility of a faithful deontic logic poses a threat to this way of defending the autonomy of ethics. If it makes sense to think of 'ought' as a sentential operator, like deontic logics do, then there will be truths of deontic logic that contain 'ought' non-vacuously and that can be deduced from no premises (and so any 'ought'-free sentences as well). Perhaps $O\phi \rightarrow \neg O \neg \phi$ (If it ought to be the case that ϕ, then it is permissible for it to be the case that ϕ) is one such example that can deduced from any premises.[10] For this reason, Pigden (1989,

[8] The difference between the class of arguments considered by the two authors to be the domain of the autonomy of ethics will be irrelevant to this discussion.

[9] But, if one does focus on those details, one should also see Nelson's (1995) critique of Pigden.

[10] Of course, any particular example is contestable, but any interesting deontic logic will admit some similar example.

p. 138-145) must reject that deontic logic is a feasible project, and he gives a series of arguments against it. This seems like a problem for his account though, since deontic logic seems to be quite productive in helping us understand normative language.

1.3 Semantic Accounts of the Autonomy of Ethics

The epistemic version of the autonomy of ethics is a claim about whether we can properly reason from just non-normative premises to a normative conclusion. Jackson (1974) and Pigden (1989) both try to save that claim from Prior's counterexamples by limiting the claim to a class of arguments in terms of the surface syntax of the arguments. When we reason though, we don't reason in terms of surface syntax. Instead, we reason about the *meanings* of claims. So, to understand the autonomy claim, I suggest that we focus our attention on the connection between the *contents* of normative and the non-normative sentences in arguments.

Russell and Restall (2010) share this intuition and attempt to understand the autonomy claim in terms of the semantics of normative language. They construct a proof of the autonomy claim in terms of Kripke models for a particular deontic logic. To do so, they must stipulate complex definitions of 'normative' and 'descriptive,' which seem not to map on exactly to the pre-theoretic notions. Their account also isn't very enlightening about what's going on in Prior's purported counterexamples.

As an alternative, I'll provide another semantics-focused account of the autonomy claim. Instead of using deontic logic, I'll employ standard semantics for normative language that assumes only that sentences are evaluated for truth with respect to, at least, an ordinary possible world and a normative standard. I'll show that this offers a much simpler way to understand and defend Hume's dictum. This new approach can also explain the deficiency in Prior's examples, unlike the other approach.

1.3.1 Russell and Restall's Approach

The semantic approach of Russell and Restall (2010) takes the autonomy claim to be an instance of a more general notion of an implication barrier. There is an implication barrier from one class of sentences to another when no sentence of the second type is implied by a collection of sentences of the first type. Restall and Russell provide a general barrier theorem that they use to derive an implication barrier from descriptive sentences to normative sentences, under a particular understanding of these classes of sentences in their models of deontic logic. The models they use are standard models consisting of a set of worlds W and a relation S supplemented with a distinguished actual world where sentences are evaluated. The relation is one of deontic accessibility, so that if all of the worlds accessible from the actual world are P-worlds, then P is obligatory. They also assume that S is transitive, euclidean, serial and secondarily reflexive, though not all of these assumptions are necessary for their proof.

The implication barrier theorem they employ works in cases where there is a set of sentences that are not preserved under extensions of the models or under changes of the worlds related by S. Normative claims appear to be such a class in the deontic models they consider. For example, while it's impermissible for Alice to hit Bob, it's permissible for Alice to hit Bob when they are in a boxing class. More formally, Russell and Restall take it that a claim is *normative* when it's truth-value is always changeable either via model extension or tinkering with S in every deontic model. A claim is *descriptive* when it is always preserved under tinkerings with S. It then follows as an instance of their Barrier Construction Theorem that no satisfiable collection of descriptive sentences entails a normative one. For a more precise statement of their result, see Russell and Restall (2010, p. 252–7).

1.3.2 A New Semantic Approach

Russell and Restall argue that we can prove the autonomy of ethics by understanding it terms of the semantics provided by models of deontic logic and a complicated constructed understanding of 'normativity' and 'descriptivity.' Here, I aim to secure the autonomy of ethics in terms of the semantics of normative language in a more straightforward way. I will assume that we interpret sentences with respect to points of evaluation that consist of (perhaps among other things) an ordinary possible world and a normative standard. By doing so, the solution remains agnostic about interpreting the semantics as relativist, contextualist, or invariantist.[11] For

[11] On the double indexing sort of semantics inherited from Stalnaker, Kaplan, and Lewis, points of evaluation are structured in terms of a context that takes expressions into contents, and an index (or circumstance of evaluation) that takes contents into extensions. According to the contextualist, the relevant normative standard figures in the context and affects what proposition is expressed. According to the relativist, the normative standard figures into the index and affects the truth value of the proposition expressed in a context. According to the invariantist, the normative standard is contained in the structure of the possible world, which is itself standardly treated as a coordinate of the index. (In order to distinguish the interpretations of normative sentences that differ in truth value, the invariantist will either need to allow impossible worlds that differ in which normative standard is contained in the world or allow that the normative aspects of worlds are not fixed by the non-normative aspects.) But what is important here is that normative sentences are evaluated for

clarity of presentation, I will use a particular modeling framework for semantics, the one given by Gibbard's *Thinking How to Live* (2003), but the result should hold for any plausible normative semantics.[12]

On Gibbard's modeling framework, sentences are modeled with sets of pairs consisting of a possible world and a complete plan for action in any hypothetically possible scenario. Then, like any standard semantics, we model the content of a sentence 'P' with the set of world-norm pairs that are compatible with that sentence. Standard possible world semantics is a special instance of Gibbard's semantics. When we have a non-normative sentence like 'Snow is white,' we can represent it as the collection of world-norm pairs where the world is one in which snow is white. The semantics of 'snow is white, or Mark ought to cook' is given by the set of world-plan pairs where either snow is white in the world or the plan calls for Mark cooking.

A sentence will be non-normative, on this kind of semantics, when it is norm-invariant, in that if some world is included in a pair in the set, then that world also appears in the set in a pair with every possible plan. More explicitly, if W is the set of possible worlds and N is the set of all complete plans,

NORM-INVARIANT a set of world-norm pairs $P = \{\langle w_i, n_j\rangle\}$ is *norm-invariant* iff $(\forall w \in W)(\exists n(\langle w, n\rangle \in P) \to (\forall n \in N)(\langle w, n\rangle \in P))$.[13]

Intuitively, a set of world-norm pairs is norm-invariant when it doesn't tell us anything normative, since it's compatible with all possible plans.

Consider some argument from premises $\{P_i\}$ to conclusion C. We might be inclined to flesh out the requirement of the autonomy of ethics on this argument like this: If each of the P_i is norm-invariant, then C is norm-invariant. This is a natural way to flesh out the idea that the autonomy claim requires that if all the premises are non-normative then the conclusion must be non-normative. But this simple suggestion fails just like SIMPLE AUTONOMY does. Consider again Prior's first example:

(A) 1. Tea-drinking is common in England.
2. Therefore, either tea-drinking is common in England or all New Zealanders ought to be shot.

truth with respect to points of evaluation that consist of (at least) a world and a normative standard. Many thanks to Alex Silk for helping me to clarify the flexibility of my proposal along these lines.

[12] Gibbard supplies this formal modeling framework to defend an expressivist account of normative language. I am only using the framework only in ways that are perpendicular to the dispute over expressivism.

[13] This is a restatement of the definition of 'norm-invariant' employed by Schroeder (2011).

The premise is norm-invariant, but the conclusion is not. The set of world-norm pairs that represents the conclusion includes a world where there is no tea and all New Zealanders ought to be shot, but it doesn't include a tea-less world where some New Zealanders ought not be shot. So, it is not norm-invariant, and this world-norm-style analogue of SIMPLE AUTONOMY fails.

That analogue of SIMPLE AUTONOMY fails because the conclusion tells us something normative. But, as Prior noticed, there seems to be something deficient about the normativity in the conclusion. When we consider this failure in the terms of semantics that evaluates sentences with regard to a possible world and a normative standard, we're better positioned to diagnose that deficiency. The conclusion of the puzzle argument only seems to give substantive normative guidance about what ought to be the case in worlds unlike the ones described by the premise, namely ones in which tea-drinking is not common in England. If we're in a world where tea-drinking is common in England, the conclusion doesn't tell us what ought to be the case here. When we consider the set of world-norm pairs compatible with the conclusion that are also compatible with the premise, we notice that they are norm-invariant. Given the premise, the conclusion tells us nothing about how things ought to be.

When Hume introduces the autonomy claim, he complains that his contemporaries are improperly reasoning from claims about how things are to how they ought to be. But, I claim, the kind of normativity present in the offending argument from Prior isn't the sort of thing Hume would complain about. It's not good practice to reason from 'Tea-drinking is common in England' to the offending conclusion not because the argument is invalid; rather, what's odd about drawing that normative conclusion from that description of England is that the normative aspect of the conclusion is irrelevant to the possibilities being reasoned about, namely ones in which tea-drinking is common in England. The argument would be something for Hume to complain about only if the conclusion made a claim about how things ought to be *in worlds where the premises are true*.

We can easily restate this intuition in our semantic framework. To decide whether the conclusion of an argument makes a claim about how things ought to be in the worlds described by the premises, we just restrict our attention to those worlds. In our semantics, when the premises are norm-invariant, deciding this question is equivalent to deciding whether the conclusion *conjoined with the premises* is norm-invariant. This results in a reformulated version of the autonomy of ethics in Gibbard's semantics:

WORLD-NORM AOE If each of $\{P_i\}$ is norm-invariant, then if $P_1 \wedge P_2 \wedge \ldots$ is satisfiable and $\{P_i\} \vdash C$, then $P_1 \wedge P_2 \wedge \ldots \wedge C$ is norm-invariant.

Intuitively, WORLD-NORM AOE tells us that if the premises to an argument are norm-invariant, then the set of all world-norm pairs compatible with the conclusion *and the premises* is also norm-invariant. By checking only the conjunction of the premises with the conclusion for norm-invariance, we restrict our attention to only those worlds where the premises are true. The satisfiability condition is included to avoid the special case where non-norm-invariant claims follow trivially from contradictory premises.

As it is presented here, WORLD-NORM AOE is clearly tailor-made to avoid the first of Prior's supposed counterexamples. WORLD-NORM AOE also avoids the other puzzle cases. Consider this one:

(C) (a) Undertakers are church officers.
(b) Therefore, if church officers ought to be reverent, undertakers ought to be reverent.

Here, the conjunction of the conclusion with the premise is a first-order tautology, so it is compatible with every world-norm pair. Since every world-norm pair is in the semantics of the conclusion conjoined with the premise, *a fortiori* every world in a pair in the set is also paired with every possible norm. So it is norm-invariant.[14]

In fact, we can show that WORLD-NORM AOE is *guaranteed* by a semantics like the one we're considering. Here's the proof: For some valid argument to C from $\{P_i\}$, suppose that each of $\{P_i\}$ is norm-invariant, $P_1 \wedge P_2 \wedge \ldots$ is satisfiable, and $\{P_i\} \vdash C$. Then since the argument is valid, C must be satisfied by every world-norm pair that satisfies $P_1 \wedge P_2 \wedge \ldots$. So, $P_1 \wedge P_2 \wedge \ldots \wedge C$ is satisfied by every world-norm pair that satisfies $P_1 \wedge P_2 \wedge \ldots$. Without loss of generality, let $\langle w, n \rangle$ be some world-norm pair that satisfies $P_1 \wedge P_2 \wedge \ldots \wedge C$. Since $\langle w, n \rangle$ must satisfy each of the $\{P_i\}$ and each of the $\{P_i\}$ is norm-invariant, for each $\{P_i\}$ $\forall n \in N(\langle w, n \rangle \in P_i)$. So $\forall n \in N(\langle w, n \rangle \in P_1 \wedge P_2 \wedge \ldots)$. Then since every world-norm pair that satisfies the conjunction of the premises must satisfy the conclusion, $\forall n \in N(\langle w, n \rangle \in P_1 \wedge P_2 \wedge \ldots \wedge C)$. So $P_1 \wedge P_2 \wedge \ldots \wedge C$ is norm-invariant.

Since WORLD-NORM AOE is a theorem, if WORLD-NORM AOE is a proper formulation of the autonomy claim (and our semantics is correct for these claims), then

[14] Prior's other supposed counterexample can be handled similarly. The conjunction of 'There is no man over 20ft high' with the conclusion that 'There is no man over 20ft high who ought to sit in an ordinary chair' is equivalent first-order equivalent to 'There is no man over 20ft high,' which is norm-invariant.

this formulation of the autonomy claim avoids Prior's counterexamples. One might worry though that it avoids Prior's counterexamples only by excluding too many arguments from the intended domain of the autonomy of ethics. One way to put the worry is like this: WORLD-NORM AOE claims that to check an argument for a violation of the autonomy of ethics, given some norm-invariant (so, non-normative) premises, we only need to check that the conclusion conjoined with those premises is also norm-invariant. But for any deductively valid argument, the conclusion conjoined with the premises is logically equivalent to the conjunction of the premises. So, WORLD-NORM AOE really only seems to check arguments from some premises to themselves, the objector claims.

The objector is right that WORLD-NORM AOE essentially limits the scope of the autonomy of ethics to arguments from premises to the conjunction of those premises, but this isn't problematic. The reason why is that if normativity arises only when the conclusion of an argument is strictly weaker than the conjunction of the premises, no reasoner employing the argument could properly come to a conclusion about what they ought to do. The world-norm semantics gives us a natural framework for understanding why this is so: Suppose Rachel wants to know whether she ought to donate to charity. We can represent Rachel's mental state by the collection of world-norm pairs M compatible with what she believes and the norms she accepts. For Rachel to become decided on whether she ought to donate, it is for the plan in each world-norm pair in M to call for donating in the (centered) possibility given by the world of the pair. If Rachel's mental state doesn't already commit her to a view on whether she ought to donate, there must be two world-norm pairs $\langle w_1, n_1 \rangle$ and $\langle w_2, n_2 \rangle$ in M such that n_1 calls for donating in w_1 and n_2 permits not donating in w_2. Now suppose Rachel happens across a valid argument of the kind the objector above is concerned about, namely one where the premises are norm-invariant but the conclusion is not (though the conclusion conjoined with the premises is). The only way for Rachel to properly become committed to donating in response to the argument is for the proper response to the argument to require that she remove one of those two world-norm pairs from the representation of her mental state. Since the argument is valid, Rachel must either accept the conclusion or reject a premise. The premises are norm-invariant, so we can think of them as ordinary centered possible world propositions. Rachel accepts them when every world in a world-norm pair in M is contained in every premise. So, if Rachel does accept them, w_1 and w_2 must be in every premise. Since the premises are norm-invariant, the premises must also have w_1 and w_2 paired with every possible norm, so importantly, $\langle w_1, n_1 \rangle$ and $\langle w_2, n_2 \rangle$ must be in every premise. But, since those world-norm pairs satisfy the premises and the argument is valid, they must also satisfy the conclusion. So when Rachel accepts the conclusion, those pairs will remain in her mental state, and she'll remain uncommitted about whether to donate. On the other hand, if Rachel does not accept the premises of the argument, her mental state will not change either. So here again, she will remain uncommitted about whether to donate.

This proof only works for deductive arguments because it requires that the argument be non-ampliative. By accepting the conclusion of such an argument, the reasoner is not committed to anything she wasn't committed to before considering the argument. Suppose though that our reasoner does become committed to the premises or conclusion of the argument even though she wasn't before. This is an odd thing for her to do, epistemologically speaking, in response to an argument, but it would make it possible that the reasoner would come to a decision about whether to donate in response to the argument. If the reasoner does become committed about whether to donate by accepting either a premise or the conclusion of the argument, it must be because one of the two world-norm pairs $\langle w_1, n_1 \rangle$ and $\langle w_2, n_2 \rangle$ is removed from M. Assuming that she doesn't learn anything strictly stronger than the premises or the conclusion, this requires that the part of the argument she comes to accept is false at either w_1 or w_2. But since the parts of the argument are all silent about what ought to be the case at w_1 and w_2 (since they include those worlds paired with every possible norm), the reasoner still doesn't learn anything normative from the argument. This means that even if the reasoner acts in what appears to be an epistemically irresponsible way in response to the argument, either by accepting a premise or the conclusion without already having been committed to it, the information that the reasoner gains from the argument can still be characterized in a norm-invariant way. So, if the reasoner does become committed about whether to donate by accepting parts of the argument, she does so only because she is already committed to a conditional of the form 'If $w_{1(/2)}$ is not actual, then I ought (not) donate.' It is this conditional, which she must already be committed to and which doesn't follow from any part of the argument, that must play the role of the normative bridge principle in her reasoning.

The motivating idea behind the above considerations is this: For arguments from non-normative premises to a normative conclusion, the normativity in the conclusion cannot be relevant to the possibilities described by the premises. Since a deductive argument could only help us learn something about how things ought to be insomuch as we accept the premises, any potential normative guidance that could be derived from non-normative premises must only really apply to possibilities where the premises fail. When we ignore those possibilities by restricting our attention only to the possibilities compatible with the premises, the normativity disappears. So, WORLD-NORM AOE is correct to effectively limit the scope of the autonomy of ethics to arguments from premises to a conjunction of the premises with the conclusion. No reasoner employing a valid argument to a weaker conclusion could properly decide what they ought to do just in virtue of accepting the argument. So any normativity present in that conclusion is deficient for reasoning about how things ought to be and can be ignored for the purposes of understanding the autonomy claim.

Understanding the autonomy claim in terms of WORLD-NORM AOE also allows us to clarify what's going on with seemingly puzzling arguments involving conceptual and analytic truths. Suppose that 'courageous' just means the same as the conjunction of 'done in the face of danger' and 'ought to be done'. Then consider this argument:

1. It's dangerous for Jeb to run into the burning building to save the cat.
2. So, if Jeb ought to save the cat, it would be courageous.

This argument is valid, but it also appears to be a case where we can derive something normative from only non-normative premises. But, we see that even though the conclusion is normative, it is norm-invariant when conjoined with the premise. This shows that, according to my formulation of the autonomy claim, this argument is not a counterexample to that claim. Intuitively, this is the right result, as a reasoner employing this argument cannot learn anything new about what ought to be the case, much like Rachel above. So WORLD-NORM AOE gets the right result about arguments like this that rely on conceptual and analytical truths.

Finally, one might worry that the understanding of the autonomy claim that I provide is too restricted because it applies only to valid, non-ampliative arguments. One might think there is a plausible version of the autonomy claim that is meant to apply to ampliative arguments. But the general version of this idea is clearly implausible, since we could have an odd ampliative logic that permits arguments from 'snow is white' to 'Rachel ought to donate'. In the case of particular ampliative logics, like logics of enumerative induction, an autonomy-like claim may hold. A full understanding of those kinds of autonomy claims would have to proceed on an individual basis for each logic though, which is not a task I will take up here.

WORLD-NORM AOE requires taking a particular stance on what is the brand of normativity that the autonomy claim is meant to apply to. It is the kind carved off by norm-invariance in semantics that evaluate sentences with respect both to a possible world and a normative standard. I presented it here using the formal modeling framework offered by Gibbard (2003) as an example, but the result will hold for any plausible semantics of normative language. As long as the semantics evaluates the truth value of a sentence at a possible world and a normative standard, WORLD-NORM AOE is a theorem. So, WORLD-NORM AOE is another semantically-driven way to understand and defend the autonomy of ethics.

1.3.3 Comparing the Semantic Approaches

The account of the autonomy of ethics supplied here and the one by Russell and Restall (2010) both aim to make sense of the claim in terms of the semantics of normative language. There are many differences between our approaches though.

Russell and Restall's approach employs Kripke models of deontic logic, and claims that the class of normative sentences are those that have a certain feature with respect to that class. They treat the class of descriptive sentences similarly. Their stipulated account of normative sentences doesn't seem to capture all of the sentences that we might pretheoretically count as normative though. Peter Vranas (2010), for example, shows that the sentence 'All citizens ought to vote' does not count as normative on their account. Certainly though, as Vranas continues, a hypothetical argument from descriptive premises to that conclusion is in the intended domain of the autonomy of ethics. This appears to be a quite problematic result for taking Russell and Restall's understanding of the autonomy of ethics to be fully general.

The reason that 'All citizens ought to vote' doesn't count as normative on their account is because the sentence is captured in their semantics as a universally generalized disjunction. One of the disjuncts of each disjunction is non-normative, namely that the object is a citizen. This results in the disjunction not being always unstable under both model extension and tinkering with the accessibility relation, which violates their definition of 'normative.' That said, their account also doesn't count the sentence as descriptive, as it is not always stable under model extension either. This shows that their account is essentially incomplete.

The semantic account of the autonomy claim that I provide does count 'All citizens ought to vote' as normative, since it is not norm-invariant: that all citizens ought to vote is incompatible with a world-norm pair where the norm calls for some citizens not to vote. So, the account I provide would block arguments to this conclusion. In fact, my account will put any non-trivial normative claim in the class of claims that we can't derive from the non-normative. This is because if the claim rules out any ways the combination of the world and norms could be, it will not be norm-invariant.[15] This completeness does come at a cost though.

A benefit of Russell and Restall's account is that it delineates two classes of claims, the descriptive and the normative, and says that there are no valid arguments from sentences in the first class to a conclusion in the second. Though not every sentence falls into one of the two classes, putting the autonomy claim in terms of classes of sentences for which there is no argument from one to the other is straightforward in a way my account is not.

My account puts a restriction on what arguments are possible, but that restriction is not statable as a restriction on arguments from one class of sentences to another. My semantic account says that there are no valid arguments from norm-invariant premises to a non-norm-invariant conjunction of a conclusion with the premises. So what class of sentences cannot be derived, on my view, is dependent on the premises being used. So although my account does seem to be fully general, it does so at the cost of capturing the autonomy claim as a restriction on arguments from one set of sentences to another.[16]

On Russell and Restall's approach, the account of the autonomy of ethics falls out of a unified understanding of implication barriers. Their account attempts to reconcile the autonomy claim with other barriers to implication, including inferences from particular claims to general claims, inferences from claims about the past to claims about the future, and inferences from claims about actuality to claims about necessity. The account I offer does not seem well-positioned to offer similar

[15] Notice though that there is still a class of claims that contain normative terms that my account doesn't count as normative, namely those that are trivially true. These include conceptual truths, like 'If you ought not sit, you're not permitted to sit' but also ordinary trivialities with normative content like 'Either you ought to sit, or it's not the case that you ought to sit.' This is a benefit of the account though, as certainly, the autonomy claim is not meant to block inferences to these kinds of claims.

[16] Notice that there is a general reason to think that no account of the autonomy of ethics could both formulate the claim as blocking arguments from one class of sentences to another and also be complete, in that it puts every sentence into exactly one of those classes. Take a general version of one of Prior's examples: '*P*. Therefore, *P* or *Q*.' where *P* is non-normative and *Q* is normative. For the account to put every sentence into exactly one class, it must count '*P* or *Q*' as normative or non-normative. If it's normative, then the given example is a counterexample to the autonomy claim. If it's non-normative, then we can generate a new counterexample: 'not-*P*. *P* or *Q*. Therefore, *Q*.' which would be a counterexample. So any account of the autonomy claim must suffer either the incompleteness disadvantage that Russell and Restall's faces or the disadvantage of not treating the autonomy of ethics as blocking inferences from one class to another.

explanations, since it'd be odd for a semantics to evaluate sentences in regard to the parameters required for the analogous results.

Despite that, my approach does defend the autonomy of ethics with only the weak assumption that our semantics models content with something like an ordinary possible world and a normative standard. This framework is highly flexible and isomorphic to many standard semantics for normative language. If Hume's dictum relies on only this weak assumption, it is much more secure than if it relies on the more controversial semantics provided by deontic logics.

Another significant benefit of my approach is that it comes along with an explanation of what goes wrong in Prior's examples and other valid arguments from non-normative premises to normative conclusions. On the epistemic reading of the autonomy claim, that claim is about whether we can properly reason from non-normative premises to a normative conclusion. According to WORLD-NORM AOE, the autonomy of ethics is not challenged by the purported counterexamples because we have independent grounds for thinking that no reasoner could properly employ those arguments to reach a conclusion about what ought to be the case in the situation they take themselves to be in. Putting the autonomy claim in terms of a barrier to implication, as Russell and Restall do, does not offer an analogous explanation.

It seemed from the outset that the autonomy of ethics was doomed by Prior's examples. Others have tried to save the claim by syntactically restricting its intended domain, like Gibbard who proposes that we only view it as a limit on arguments to simple normative conclusions. But this move is both too strong and too weak. We ought instead to seek a semantically-motivated understanding of the autonomy claim.

Russell and Restall's approach, which is grounded in deontic logic, offers one way to do that, but, as Vranas shows, it seems to be faced with a completeness worry, like the syntactic approaches. I've introduced a new way to understand and defend the autonomy of ethics in terms of a more general semantics of normative terms. The new version seems to offer the required flexibility to make sense of the autonomy claim, and it also offers a nice story about what goes wrong in Prior's cases. It does this at the cost of not being able to treat the autonomy claim an instance of the unified notion of an implication barrier. On my account, whether a conclusion is in the class of sentences that the autonomy of ethics is meant to block is relative to the particular premises in play: it says we can't get from non-normative premises to conclusions that are normative when the premises are true.

CHAPTER II

On a Puzzle about Correct Belief

ABSTRACT

From "Snow is white," we can infer "If Dan believes snow is white, his belief is correct." Saying that a belief is correct is to say that it is the belief one ought to have, for some sense of 'ought.' But this inference seems to violate Hume's autonomy of ethics, the claim that what *is* the case doesn't fix what *ought to be* the case. Gibbard (2012, Ch. 4) proposes a solution to this puzzle. His proposal tries to resolve the puzzle by trying to understand the kind of normativity that occurs in it. Gibbard claims that by understanding how objective and subjective oughts relate to each other, we'll see that the offending argument doesn't really bridge the is-ought gap.

I cast doubt on Gibbard's proposed solution by claiming the no epistemology of the normative fits nicely with it. Any epistemology that Gibbard could supply would either be unable to explain an awkward asymmetry it must posit in the subjective ought or make subjective oughts too detached from the agents they apply to. In arguing for this disjunction, I also argue that the subjective ought does not play the role in our deliberation that Gibbard expects it to. Really, there is no unique ought that factors in our deliberation.

We should reject Gibbard's proposed solution to the puzzle, I claim, because it gets wrong the relationship between objective and subjective oughts. I suggest another solution to the puzzle, one that takes the concept BELIEF to be normative.

Hume taught us that no 'ought' follows just from an 'is'. That is, no substantive normative claims are fixed by non-normative facts. This is the putative *autonomy* *of ethics* or the *is-ought gap*. There is a seemingly valid argument regarding correct belief that appears to violate the autonomy of ethics and is discussed by Gibbard (2012, Ch. 4). Gibbard proposes a way to make sense of the puzzle while maintaining the autonomy of ethics. He claims that the conclusion of the offending argument is trivial in a way that doesn't threaten Hume's dictum. I agree with his verdict but disagree with Gibbard as to why. In this chapter, I'll lay out the puzzle and Gibbard's proposed solution. I'll show why one ought to reject Gibbard's response to the puzzle and show how the type of view advocated by Shah (2003) can also explain the puzzle. In the next chapter, I provide a new argument for the key premise of that second solution.

2.1 The Autonomy of Ethics

In the *Treatise of Human Nature*, Hume states the autonomy of ethics. He says:

> In every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary ways of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when all of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not. This change is imperceptible; but is however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given; for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it. (Hume 1978, p. 468)

Inferring normative facts from only non-normative facts is unwarranted, Hume claims, because "this new relation" of normativity seems "entirely different" from descriptive or natural propositions. What *is* the case, the ways things *are*, does not seem fix at all how things *ought* to be. Here is a first pass at formalizing the autonomy of ethics:

SIMPLE AUTONOMY The normative isn't determined by anything non-normative.

SIMPLE AUTONOMY, if right, would have it that no way of fixing the descriptive facts is enough to figure out how things ought to be. It follows from this that we can't properly argue from how things are to how they ought to be – that is, there is no valid argument with only descriptive premises and a non-trivial normative conclusion.[1]

A. N. Prior (1960a) showed us that SIMPLE AUTONOMY fails. Here is one of his three examples:[2]

(1) Tea-drinking is common in England.
∴ (2) Tea-drinking is common in England or all New Zealanders ought to be shot.

Like this one, each of Prior's examples is a logically valid argument with non-normative premises and a conclusion that contains a normative term essentially.[3]

Many ways of resurrecting the autonomy of ethics in light of these counterexamples and others have been proposed. See, for example, Jackson (1974), Pigden (1989), and Gibbard (2012, p. 80-81). Each of these accounts tries to save the autonomy of ethics syntatically – that is, by restricting the class of normative claims that the autonomy of ethics is meant to apply to by roping off some class on the basis of its syntax. In Chapter I, I suggest that roping off the claims solely based on the syntax doesn't best capture the idea of the autonomy of ethics. Instead, we need a semantic account of the autonomy of ethics. I provide such a semantic account and argue for it there. Though the details of the argument here won't depend on accepting *my* account of the autonomy of ethics, they will require accepting some account that avoids Prior's counterexamples.

2.2 The Puzzle about Correct Belief

The purported counterexample to THE AUTONOMY OF ETHICS that Gibbard (2012, Ch. 4) considers is this: If Sophia believes that snow is white and snow actually is white, then her belief is correct. The judgment that Sophia's belief is correct seems to be a normative judgment. It's the judgment that Sophia's belief is the belief she ought, in some sense, to have. Further, that normative claim seems to follow from the non-normative claim that the content of the belief is true. This appears to violate the Humean dictum discussed above that normative conclusions can't follow from merely non-normative premises.

[1] Here, I'm assuming that the normative conclusion is dependant on the premises and hence not derivable from the empty set.

[2] Here I actually provide the slightly modified version from Pigden (1989, p. 132).

[3] A term is *essential* when it cannot in general be replaced with another term of the same grammatical category without a change in truth-value.

More explicitly, the argument of the puzzle goes like this:[4]

(1) Snow is white.
(2) Sophia believes that snow is white.
∴ (3) Sophia's belief that snow is white is correct.

The autonomy of ethics is supposed to block valid inferences of a normative claim from non-normative premises. So the puzzle is one of how this argument about belief correctness can be made to cohere with the autonomy of ethics.

What is interesting about this particular puzzle is that the argument doesn't seem to be like Prior's examples. In each of Prior-style examples, satisfying the premise makes it trivially true that the normative conclusion holds, so that conclusion is no longer potentially guiding or prescriptive of action.[5] Given that tea-drinking is common in England, that tea-drinking is common in England or all New Zealanders ought to be shot comes for free. But, the correctness of Sophia's belief doesn't appear to be trivial in the same way. In the end, though the correctness conclusion doesn't seem to be trivial, I will argue that it really is.

2.2.1 *Three Ways to Defuse the Puzzle*

In broad brush strokes, there can be three kinds of approaches to this puzzle that maintain both the autonomy of ethics and the truism that a belief is correct if its content is true.

The first kind of response to the puzzle holds that the argument is strictly speaking invalid. Proponents of this view might suggest that the argument *seems* valid because we're typically willing to grant an implicit normative premise – the premise that a belief is correct if its content is true. According to this response, the puzzle is resolved because the purported counterexample is not valid without the hidden normative premise. I'll call this kind of response an *implicit premise response*

[4] My formulation of the argument differs slightly from Gibbard's in that Gibbard's doesn't contain the second premise. The conclusion of his puzzle is that Sophia's belief that snow is white is correct. His formulation of the argument is enthymematic: it requires that Sophia exist in the first place and that Sophia has the relevant belief. Boghossian (2003, p. 37) also recognizes the second requirement. Strictly speaking, the conclusion would not obtain were one of these to fail, and the argument is invalid without these. Let's grant that Gibbard would include these premises since the mere addition of the premises is not enough to solve the puzzle.

[5] My formulation of the autonomy of ethics allows us to make more sense of this notion of triviality. More precisely, though the conclusion of the argument is normative, it cannot potentially guide a reasoner to a normative conclusion she did not already accept. See Chapter I.

to the puzzle.[6] This kind of approach is investigated by Horwich (1990), (2000), and Gibbard (2012, p. 80).

A second kind of response attempts to resolve the puzzle by denying that the conclusion of the offending argument is normative. According to this kind of response, correctness of belief, or at least the brand of correctness of belief employed in the conclusion of the puzzle, is a non-normative notion, conveying not much more than the truth (conceived of non-normatively) of the belief's content. I'll call this kind of response a *non-normative conclusion response*.

The final kind of response doesn't deny the normativity of correctness, but rather it denies that the conclusion of the purported counterexample is a non-trivial normative claim given the premises. That a belief that P is correct given P, on this view, is a trivial normative claim like the conclusions of Prior's examples. Deriving insubstantial or non-trivial normative conclusions from non-normative premises doesn't challenge the autonomy of ethics because, in a sense, these claims don't really tell us anything about how things ought to be apart from how things are.[7] I'll call this kind of response a *trivial conclusion response* to the puzzle. Gibbard's solution is one such response. Shah (2003) and Shah and Velleman (2005) argue that the concept BELIEF is a normative one. On this proposal, it's a conceptual truth that a belief is correct iff its content is true. This proposal, like Gibbard's, would also generate a trivial conclusion response to the puzzle – one that I'll endorse.[8]

2.3 *Gibbard's Trivial Conclusion Response*

The response that Gibbard proposes in chapter 4 of *Meaning and Normativity* is a trivial conclusion response. First, Gibbard tells us that *correct belief*, which factors in the conclusion of the puzzle, is the belief that one *ought* to have, for some appropriately-conceived objective sense of 'ought' (2012, p. 75).[9] Gibbard's plan is to claim that the triviality of the conclusion drops out of a proper account of this sense of 'ought'.

To make good on this, Gibbard gives a series of characteristics the suitable sense

[6] Thanks to J. Dmitri Gallow for first suggesting this kind of response to me.

[7] Any account that avoids Prior's worries for the is-ought gap will say this. See Chapter I for further discussion of one way to understand it.

[8] This solution will also say that the argument escapes the autonomy of ethics worry because the second premise is normative. None-the-less, the conclusion is still trivial given the normative premise.

[9] I take the move here to be an endorsement of Ewing's (1939) and (1955, p. 341-342) proposal that all normative claims are synonymous with ought claims – i.e. that ought is the fundamental normative concept. The earlier Gibbard (1990, p. 42) endorses this move.

of ‘ought’ must have. Since Sophia ought, in this sense, to believe that P when P is true, regardless of the situation Sophia is in, the ‘ought’ must be one that ignores costs and delivers ideal standards (2012, p. 78). Importantly, the norm generated is also supposed to hold even of agents who don’t have all of the facts accessible to them, as it is independent of the agent’s evidential state that if her belief is true, it is correct. So, the ought must be an objective one – one that applies in light of all the facts, not a subjective one – one that applies in light of the information available to the agent. The objective and subjective ought differ in that “the basic normative precepts that ground a subjective ought are subjectively applicable – applicable in light of information the agent has” (Gibbard 2012, p. 76), whereas the precepts that ground objective oughts may require information the agent has no way of accessing.

Gibbard’s goal is to give a general strategy for accounting for the meaning of objective oughts. It will fall out the meaning of the objective ought, he claims, that the conclusion of the puzzle is trivial.[10]

To begin the project, Gibbard first argues that subjective oughts cannot be reduced to objective ones – that is, the meaning of subjective ought claims cannot be accounted for in terms of objective oughts. Roughly, the idea here is that subjective ought claims contain strictly more information than objective ones. Gibbard gives this example: Suppose you’re offered a bet on a coin where you’d receive $1000 if it lands heads and lose $800 otherwise. In order to figure out what you *objectively* ought to do to maximize money gain, we’d only need to know how the coin will land and whether receiving $1000 is better or worse than losing $800. In order to figure out what you ought to do *subjectively*, we’d need to know more, namely how to compare the values of the possible money gains and loses. According to Gibbard, “In terms of classical decision theory, the utility needed for objective oughts is ordinal, whereas that for subjective oughts is cardinal: the scale must allow for comparisons of preferability in degrees” (2012, p. 81). But objective oughts don’t give us enough information to compare in degrees, Gibbard thinks. So, subjective oughts cannot be defined solely in terms of objective ones.

In the particular case of belief, Gibbard notes that the objective ought claims

[10] It may seem strange that such a project is possible. That strangeness is ameliorated when we notice that the we’re looking into doxastic, objective oughts – that is, claims about what one ought to believe in light of all information. But plausibly, in light of full information, many things that may seem strange are true. For instance, it may be that in science one objectively ought not even consider the most plausible hypothesis on one’s evidence in light of all the facts since considering whether $\neg P$ is true is silly in light of P.

about belief fix fully what one ought to believe given all the information, but this is not enough to fix what one ought to believe given a lack of information. "Much more is needed to settle what degree of credence one ought subjectively to have," Gibbard contends. Objective oughts cannot provide the needed information because were they able to, they'd answer a normative question about how to assess evidence, but it seems that merely having the facts can't answer this question. As Gibbard puts it, "A person who knows all the facts still can't determine, from this and sheer analytic definitions alone, how, subjectively, I ought to assess my evidence." (2012, p. 82).

Given the failure of giving the meaning of subjective oughts in terms of objective ones, Gibbard turns to providing the analysis in the other direction. We'll see that if Gibbard has provided the correct analysis in this direction, it will fall out of the meaning of the objective ought that one objectively ought to believe that *P* when *P*. One could give a rough gloss of Gibbard's proposal like this: *S* objectively ought to ϕ iff if *S* had full information, it would be the case that *S* subjectively ought to ϕ (2012, p. 82).

We need a few revisions to flesh out the details, according to Gibbard: First, to have full information is to believe all that's true, including the properly *de se* truths, such as those about one's perspective and ignorance. But mere believing isn't sufficient since one ought not act on mere belief. One ought act on information only if one is justified in that information (2012, p. 83).[11] Further, actually having the belief doesn't seem to matter, so long as one ought to have the belief, Gibbard claims. Hence, the proposal is refined to this (Gibbard 2012, p. 84):

GIBBARD'S ANALYSIS *S* objectively ought to ϕ iff were it that [*S*] [subjectively] ought to accept all that's so, [it would be the case that *S*] [subjectively] ought to [ϕ].

If this proposal about the meaning of the objective ought is correct, then the puzzle about belief is resolved. The strategy is this: Take any true proposition *P*. Then were it that *S* subjectively ought to accept all that's so, it would be the case that *S* ought to accept that *P*. But that is just to say that *S* objectively ought to accept that *P*, according to the proposed analysis. We can then cast a belief as correct if one objectively ought to believe it. From this, it falls out that a belief that *P* is correct, when *P*. So the normative conclusion that follows from the non-normative premises

[11] Others think the condition on proper action is strictly stronger, such as Hawthorne and Stanley (2008) who think knowledge is required. Others might think Gibbard is mistaken and that mere belief is all that is required.

is trivial – it falls out of the meaning of the objective ought. Since any account of the autonomy of ethics that avoids Prior's worries will hold that trivial conclusions aren't in the domain of the autonomy claim, the puzzle is no longer threatening.

Put another way, Gibbard proposes that "Bob's belief that snow is white is correct" is synonymous with "Bob objectively ought to believe snow is white", which is in turn synonymous with "Were it that Bob subjectively ought to accept all that's so, it would be the case that Bob subjectively ought to accept that snow is white." Then we can transform the puzzle's argument by substitution of synonyms to an argument with the premise "Snow is white" and whose conclusion is "Were it that Bob subjectively ought to accept all that's so, it would be the case that Bob subjectively ought to accept that snow is white." This conclusion is clearly trivial given the premise. So, the original argument isn't a counterexample to the autonomy of ethics. That is Gibbard's resolution to puzzle about belief in *Meaning and Normativity*.

2.4 On Subjective and Objective Oughts

The difference between the objective and subjective ought is often cast in terms of what information the ought applies in light of. I'm using 'information' here to just mean a collection of propositions. We'll want to view these propositions as finely individuated, so that two normative sentences may pick out different propositions even if they share a truth value in every possible world.[12] Here, I'll mostly appeal to natural intuitions about 'in light of' sentences. We can make sense of what one ought to do in light of being a father, and we can contrast this with what one ought to do in light of being a judge, for example.

The big picture of the oughts in play is this: The objective ought applies in light of everything that's true (including the *de se* truths of the agent). For example, if *S* is in a rush and is approaching a blind intersection, in light of the fact that there is no oncoming traffic, she *objectively ought* to continue through without stopping. But, in light of the information available to the subject, *S ought* to slow down and look. This second sense of 'ought' is the subjective one. What *S* subjectively ought to do is what one ought to do in light of how things appear in *S*'s subjective situation. Following Gibbard, I will also say that the subjective ought applies 'in light of the information available' in *S*'s subjective situation, where this is meant to be

[12] For example, 'stealing is wrong' and 'murder is wrong' are both necessarily true, many people believe. None-the-less, we can think of these sentences as picking out different propositions, perhaps corresponding to the semantic content of the sentence.

synonymous with 'in light of how things appear'. 'In light of the information available' has a factive reading, where some information can only be available if it's true. Neither Gibbard nor I intend to employ this reading.

In contrasting the subjective and objective oughts, Gibbard tells us:

> Now in a way, objective oughts seem fishy. It would be nice, to be sure, to know what objectively you ought to do. If indeed you could check with an omniscient advisor, that's what you would ask: If, say, you wonder whether to take an umbrella, who better to ask than someone with detailed foresight of whether it will be raining at moments you'll want to be outside? Such advisors, though, are in short supply. Your real questions, then, are what to do on the basis of information you have. . . . The ought-precepts you need, we can argue, will be for 'ought' not in the fishy objective sense, but in the subjective sense . . . (2005, p. 343)[13]

We see here that Gibbard takes the subjective ought to be the one we deliberate about. When I'm asking myself what ought I do, I'm asking myself, according to Gibbard about what I subjectively ought to do. I'll use the term "deliberative ought" to denote the ought that factors in our deliberations as Gibbard describes. Gibbard's claim then is that the deliberative ought is the subjective ought.

I'll show that we should reject Gibbard's proposed analysis of the objective ought because it doesn't sit well with any epistemology of the normative. To do this, I'll break accounts of the epistemology of the normative into two camps: According to the epistemologies in the first camp, we learn about the normative just like we learn about the non-normative – much of it is essentially gained through experience. On these views, to decide what to think about a normative proposition, we should weigh the evidence we have for and against it. As such, different fully rational people may disagree by having different evidence about the normative, which they may weigh differently for and against various normative propositions. The defining characteristic of this class of epistemologies will be that they admit the possibility that two ideally rational people can have disagreeing views about a distinctly normative proposition.[14] Some epistemologies will make normative learning *a*

[13] This article by Gibbard appears to be a precursor to his chapter 4 of *Meaning and Normativity*. I use some passages from that piece when they seem to more clearly or explicitly give Gibbard's view but not when they contradict that view.

[14] By "distinctly normative proposition", I mean a normative proposition that does not rely on any non-normative facts for its truth or falsity. For example, *if there's candy in a store, it's wrong to steal it* is distinctly normative because once we fix the normative facts, the truth or falsity of this proposition is fixed.

priori, so unlike how we learn about most of the paradigmatically non-normative, but still admit of the hallmark of this camp. Those views will be considered to be in this camp of views as well, which I'll call 'rational disagreement permitting' views.

The second camp of epistemologies of the normative say that what we know about the normative must be *a priori*. No experience of how things are directly informs us on how things ought to be, they might claim. As such, everything we know about how things ought to be must be independent of our particular experiences. Further, since fully rational agents have full access to what is *a priori*, ideally rational people would converge on the same views of normative, according to these views, just as ideally rational people would converge on views of the mathematical. Gibbard's (1990, 2003) view is like this,[15] and many ideal response theories, like those of Firth (1952) and Smith (1994, p. 187) are like this as well. These views deny that fully rational agents would disagree about the distinctly normative, so I'll call them 'rational agreement requiring' views.

I'll argue that if a view is in the first camp, the rational disagreement permitting views, and adopts GIBBARD'S ANALYSIS, then the view must posit a problematic asymmetry between the way that the subjective ought treats normative and non-normative information. On the assumption that the deliberative ought is the subjective ought, I'll provide three reasons for thinking that this ought doesn't take the normative/non-normative difference into account, as GIBBARD'S ANALYSIS would suggest. I will contend that the difference between the normative and non-normative does not, in fact, play an important role in the subjective ought. What one subjectively ought to do, I claim, is what one ought to do in light of the information available, regardless of whether is it normative or not.

I'll then argue that there isn't a unique ought that factors in our deliberations, and as such, the subjective ought is not the deliberative ought. I'll use the three reasons given to conclude that neither the subjective ought nor any ought about which we deliberate applies in light of information that is asymmetric between the normative and non-normative. But, I'll show that even without this problematic assumption, the three reasons are still problematic for rational disagreement permitting views

[15] Gibbard (1990, Ch. 3) gives an account of 'rational' whereby calling something rational is tantamount to calling it the thing to do, pace Brandt (1979, p. 11, 149). Gibbard (2003, p. 12) also infers from someone being fully rational that "she will form correct fundamental beliefs as to what she ought to do". Moreover, Gibbard (2008b, p. 10) claims that "if ought judgments are a priori, false ones are rationally ruled out." A confusing claim that Gibbard makes is that there can be ideally coherent agents who are mistaken about some normative facts. See, for example, his discussion of the ideally coherent anorexic (1990, p. 172) and the ideally coherent Caligula (1990, p. 196). As such, Gibbard takes it that being ideally *coherent* is not sufficient for being ideally *rational*.

that adopt Gibbard's solution to the puzzle of correct belief.

Gibbard's actual normative epistemology and others escape those worry by being rational agreement requiring views and requiring that the correct normative view is always available to ideally rational agents. While this move avoids the problems for the first camp, it comes at another cost. By adopting Gibbard's solution to the puzzle, rational agreement requiring views force the subjective ought to be too disconnected from the subjective situation of the agent – so much so that it doesn't even seem compatible with the triviality that what *S* subjectively ought to do is what one ought to do in light of what's available in *S*'s subjective situation. I conclude that since no epistemology of the normative – that is, no rational disagreement permitting or rational agreement requiring view – that adopts Gibbard's solution to the puzzle is acceptable, we ought to reject Gibbard's solution.

2.4.1 The Asymmetry in Gibbard's Subjective Ought

Above, I roughly characterized the subjective ought as applying in light of the information available to the agent. Gibbard's proposed understanding of the objective ought requires us to understand this in an odd way, I'll claim. If Gibbard is right, subjective oughts apply in light of only the *non-normative* information available to the subject, not in light of all of the information available to her. Gibbard seems to recognize this, saying that "The distinction [between the objective and subjective ought] is a matter of how much by way of the *non-normative* facts one needs to know to ground a non-ultimate ought" (2012, my emphasis, p. 76).

To see why this is, recall that GIBBARD'S ANALYSIS has us understand what an agent objectively ought to do in terms of what what the agent subjectively ought to in the hypothetical scenario where the agent subjectively ought to accept all that's so. To accept all that's so, Gibbard tells us, the agent "must believe a full and true factual description of [her] circumstances." (2012, p. 83). Gibbard is clear here, and in the rest of the chapter, that he is using 'factual' in a way that precludes that the term applies to normative truths. So, what one objectively ought to do is what it would be the case that one subjectively ought to do were it that one subjectively ought to accept all of the non-normative truths about one's situation. With this understanding of Gibbard's proposal, we can show that treating the subjective ought as applying in light of only the non-normative information available to the agent, rather than *all* of the information available to the agent, is essential to the kind of account Gibbard offers. Without this asymmetry in the subjective ought, GIBBARD'S ANALYSIS is implausible:

Suppose a proponent of GIBBARD'S ANALYSIS were to allow the subjective ought to apply in light of all of the information available to the agent, including the normative information. Recall that, in the sense Gibbard and I are using it, the 'information available' to the agent is roughly what the agent has access to, or how things appear to be from the agent's perspective, regardless of whether that's accurate. So suppose further that the agent accepts (and perhaps even has good reason to accept) some false normative claim, such as that needlessly inflicting pain is permissible. Of course, objectively, the agent ought not needlessly inflict pain. On this version of Gibbard's view, that objective ought claim amounts to the claim that were it the case that the agent subjectively ought to accept all that's (non-normatively) so, it would be the case that she subjectively ought not needlessly inflect pain. By stipulation, the agent actually takes it that she may needlessly inflict pain. So even if she *ought* to accept a full and accurate picture of the non-normative facts about herself, we'd still expect to her to think that needlessly inflicting pain is permissible.[16] In that counterpossibility then, in light of the information available to the agent, she may needlessly inflict pain.[17] So it is not true in this hypothetical scenario that she subjectively ought not needlessly inflict pain, as GIBBARD'S ANALYSIS would require.[18]

We see then that GIBBARD'S ANALYSIS is committed to an asymmetry in the way that the subjective ought handles normative and non-normative information in the following sense: according to the analysis, the subjective ought applies in light of the non-normative information available to the subject and the true normative information, regardless of whether the agent has access to it.

[16] I'm assuming that we evaluate this kind of counterfactual by considering the possibility where minimal changes from actuality are made to accommodate the supposition. In such a possibility, we have no reason to think the agent's normative attitudes have changed.

[17] Here I'm assuming that if an 'ought' applies in light of some normative information, then generally-speaking, the information obtains when the operator applies. More explicitly, if O is the 'ought' operator, I'm assuming:

IN-LIGHT-OF REPETITION For all P and Σ, if $O(P) \in \Sigma$, then in light of Σ, $O(P)$.

Notice that this is a very weak assumption as it just allows the repetition of some claim in Σ. It is much weaker than monotonicity of 'in light of', in the sense that if in light of Σ, P, then in light of $\Sigma \cup Q$, P. This monotonicity claim is much less plausible.

[18] Notice that a similar conclusion can be reached without the assumption that we consider a scenario in which the agent subjectively ought to accept a full and accurate picture of the just the non-normative information. Even if the agent subjectively *ought* to accept a complete and accurate picture of her normative situation, the agent doesn't necessarily do so. If the agent actually believes a normative falsehood, then the nearest possibility in which she ought not believe it is plausibly one in which she still does.

2.4.2 *The Four Envelope Problem and the Many Oughts*

With that in mind, consider this case, a variant of the Three Envelope Problem[19]:

THE FOUR ENVELOPE PROBLEM S is on a game show and must choose exactly one of four envelopes, $envelope_A$, $envelope_B$, $envelope_C$, and $envelope_D$. There are four cards, $card_0$, $card_1$, $card_2$, and $card_3$, distributed in the envelopes. If the envelope containing $card_n$ is chosen, exactly n infants, who would have otherwise died a pain- and anxiety-free death, will be saved. $card_2$ is in $envelope_C$ and $card_1$ is in $envelope_D$, and S knows that. One of $card_0$ and $card_3$ is in each of envelopes $envelope_A$ and $envelope_B$. S knows that, but he doesn't have any reason to think one arrangement is more likely than the other.

About this scenario, we can ask, 'Which envelope ought S choose?' In doing so, there are four senses of 'ought' that could be in play. For each envelope, there is a sense of 'ought' under which S ought to take that envelope.

Suppose that, in fact, unbeknownst to S, $card_3$ is in $envelope_A$. Since saving three infants is better than saving none, just one, or just two, there is a sense in which S ought to choose $envelope_A$. That is, given all the facts and given that $card_3$ causes the best possible outcome, S *objectively ought* to choose $envelope_A$.

Of course, S doesn't have enough information to know that he objectively ought to choose $envelope_A$. All he knows is that $card_3$ is either in $envelope_A$ or $envelope_B$. So, using the information S has, he can only tell that the expected values of choosing each card. So, since saving two infants is more than half as good as saving three, the expected value of S choosing $envelope_C$ is higher than the expected value of choosing either $envelope_A$, $envelope_B$, or $envelope_D$. So, there is also a sense in which S *ought* to choose $envelope_C$. This sense of 'ought' is Gibbard's subjective ought.

Now suppose, unlike me, S has been persuaded by anti-natalist arguments to the conclusion that reproduction and bringing up infants causes overpopulation, famine, and depletion of the resources required for our survival. Because of this, S believes that, all other things being equal, we ought not prevent an infant from dying a pain- and anxiety-free death. From S's subjective situation, it appears that letting infants painlessly perish is better than using our resources to raise them. So, by S's lights, choosing the envelope containing $card_0$ causes the best possible outcome. Given the fact that $card_0$ is in $envelope_B$, from S's view of things, the best

[19] This case is discussed by many authors, most notably Parfit (2011) who attributes it to Regan (1980). Wedgwood (2007a) and Kolodny and MacFarlane (2010) have analogous examples.

thing to do is choose $envelope_B$. If S is to best pursue what he thinks is the best thing to do, he ought to choose $envelope_B$. So, there is a also sense of 'ought' in which S ought to choose $envelope_B$.

Finally, because the amount of resources required to save the infants varies proportionally to the number of infants saved, if the anti-natalist arguments are right, then the value of saving infants is inversely proportional to the number of infants saved. Then by S's lights, it is much worse to get $card_3$ than it is to get $card_1$. For S to rationally pursue the optimal outcome according to his own beliefs and values, S ought to choose envelope $envelope_D$. So, there is a sense of 'ought' in which S ought to choose envelope $envelope_D$.

The many senses of 'ought' in this case are generated because different 'ought' claims apply in light of different information.[20] This general point was admitted from the outset of this discussion: the proposal under consideration from Gibbard requires that objective oughts apply in light of all that so, whereas his subjective oughts apply in light of the non-normative information available to the agent. So, Gibbard admits that for different collections of non-normative information, different oughts apply. What Gibbard leaves unrecognized is that different oughts apply in light of different *normative* information as well. The agent ought to choose $envelope_A$ or $envelope_C$, but in light of S's anti-natalist commitments, consistency requires that he ought to choose $envelope_B$ or $envelope_D$.

2.4.3 *An Asymmetry in the Subjective Ought?*

In the example above, there were two collections of non-normative information in play, the facts and what the agent takes the facts to be. There were also two normative views in play, the facts about what the agent ought to do and what the agent takes to be the facts about what she ought to do. For each combination of normative and non-normative considerations, there is an ought that applies in light of those considerations.

In the order that they appear in the example above, we see the following four types of oughts:

[20] This is by no means a new point. Linguists and philosophers have long recognized that oughts come in different flavors. See for standard references Kratzer (1977), Kratzer (1981), and Kratzer (1991). Kolodny and MacFarlane (2010, Sec. 2) take Kratzer's context-sensitive approach to be importantly different in kind from Gibbard's approach, that of distinguishing objective and subjective oughts. But, whether the flavor of the ought is determined by context or is a more built-in facet of the ought, I think, is irrelevant to the discussion here.

‘Ought’s in THE FOUR ENVELOPE PROBLEM		
Type of Ought	**Non-normative**	**Normative**
Objective	All Truths	All Truths
Gibbard’s Subjective	Agent’s View	All Truths
Friendly Advisory	All Truths	Agent’s View
Subjective	Agent’s View	Agent’s View

The table lists the information in light of which the ought applies. For instance, Gibbard’s subjective ought applies in light of the non-normative facts by the agent’s lights and all of the normative truths, independent of whether the agent has access to them. The names of the types of oughts will be useful for discussion of them below. Gibbard and I both use the name “objective” for the first type, and I’ll suggest that the final type is what we typically think of as the subjective ought. I’ll call the third kind of ought “friendly advisory” because it is the ought that one would use to advise someone on what to do while granting their normative judgments.[21]

In the table, I use ‘Agent’s View’ to refer to the agent’s representation of how things are and how they ought to be. Real agents’ representations differ from the facts in two important ways though. Unlike the facts about how things are and how they ought to be, our attitude towards these things are regularly incomplete and degreed. So, in general, there is no collection of everyday propositions that fully represents the agent’s views. Rather, there are degreed and incomplete attitudes about how the world is and how it ought to be, and the subjective ought applies in light of these. In this discussion, I won’t appeal to anything that hangs on this difference between two collections of so-called ‘information.’[22]

Nothing I’ve said so far poses a problem for Gibbard’s solution to the puzzle of correct belief. We can simply view GIBBARD’S ANALYSIS as giving a reduction of the first kind of ought, the objective ought, to the second type, Gibbard’s subjective ought.

According to Gibbard though, the ought that we deliberate about treats the normative differently than it treats the non-normative. The deliberative ought, which is the subjective ought on Gibbard’s view, applies in light of the agent’s

[21] This isn’t a point that is relevant to my main claim here; the role of the name is just to make discussion easier.

[22] Notice that there is no special problem here for how an ought can apply in light of degreed or incomplete information, since there isn’t even a story about how an ought can apply in light of any kind of information yet. A full account of oughts would have to explain how oughts can apply in light of the things they apply in light of, but that task is beyond the scope of this discussion.

attitudes about non-normative matters but not in light of the agent's views on normative matters. I'll claim that, when we make decisions and deliberate about what to do, it is not so obvious that we respect or ought to respect this asymmetry.

Gibbard tells us that subjective oughts differ from objective ones in that subjective ones "are the oughts that exert normative governance – the oughts we accept and whose acceptance is directly motivating. We must act, after all, in light of information we have" (2012, p. 80). Surely Gibbard is right that we must act in light of the information that we have, but Gibbard's subjective ought doesn't respond to all of the information that we have. Agents often have normative information, and this information is left out of the picture in Gibbard's account of the subjective ought. Gibbard must explain why the subjective ought applies in light of the considerations available to us *unless those considerations happen to be normative*. Here are three reasons to expect that the asymmetry can't be explained:

First, look at standard accounts of decision theory: Decision theories are accounts of ideal deliberation, and they treat the non-normative and the normative symmetrically. What one ought to decide according to standard expected utility theory is a function of one's credences and the values one assigns to the possible outcomes – the stand-ins for one's non-normative and normative attitudes. Jeffrey (1983, p. 5), for example, fleshes out what one ought to decide in terms of "the probabilities *that the agent attributes* to the . . . conditions" and "the desirabilities *that the agent attributes* to the . . . conditions" – not, as the asymmetry in Gibbard's solution would predict, in terms of the probabilities the agent attributes to the conditions and the desirabilities *simpliciter* of the conditions. Similarly for other versions of decision theory.[23] As accounts of how we ought to make decisions, decision theories aim to give an account of proper deliberation. As such, they should be read as giving an account of the deliberative ought – one that is univocally symmetric with respect to the non-normative and the normative.

Suppose that standard decision theories are right that, ideally, the deliberative ought is in light of the agent's normative and non-normative information. Even so, that doesn't entail that, in fact, the deliberative ought is symmetric with regard to these two types of information. It could be that, in fact, when we deliberate, we deliberate about an ought that applies in light of our own non-normative information but the true norms. So this consideration regarding decision theory

[23] See for example Ramsey (1926, p. 176) who casts the desirabilities in terms of what can be solicited *from the agent* and Lewis (1981, p. 6) who casts the value of an option W in terms of "how satisfactory it seems to the agent for W to be the actual world".

doesn't give us a strong reason to think that Gibbard is wrong about there being an asymmetry in the deliberative ought. Rather, we should think that because decision theories contain the symmetry and *because it's not a surprising (or even often argued for) aspect of the theory*, that's some reason to suppose that the asymmetry doesn't exist in everyday (non-ideal) decision making. This is some reason to think that the asymmetry that Gibbard posits in the subjective ought, which he equates with the deliberative ought, is not there.

The second reason to think that the subjective ought is not asymmetric as Gibbard requires is that when deliberating or deciding what to do, we don't privilege our own non-normative information over our normative views, as Gibbard ought to predict. Consider Jaina, a young woman at the grocery store who must decide whether to buy meat. For her, the question hinges on two things:

CREDIT whether she has a sufficient amount of credit left to buy the meat, and

PERMISSIBLE whether it is morally permissible for her to eat meat.

It seems that these two propositions are on par for Jaina in the following sense: Jaina's decision depends on her views (rather than the truth) about CREDIT iff her decision depends on her views (rather than the truth) about PERMISSIBLE.

Suppose Jaina decides to buy the meat and that this was the wrong decision. Consider the following two possible excuses for Jaina's poor decision:

CREDIT EXCUSE Jaina made the wrong decision about whether to buy the meat because she didn't have enough credit. But it's ok because she had no way of knowing that.

PERMISSIBLE EXCUSE Jaina made the wrong decision on whether to buy the meat because eating meat is impermissible. But it's ok because she had no way of knowing that.

These are both accessibility-based excuses – that is, they attempt to excuse Jaina's decision because some information that ought to have factored in her decision wasn't available to her. My intuitions flip back and forth about whether either of these excuses could apply to Jaina's bad decision, but they are univocal that one rebuking of Jaina is acceptable iff the other is. If not having access to some bit of information required to make her decision is an acceptable excuse, it's acceptable excuse regardless of whether the bit of information is normative or not.

Gibbard's analysis of the deliberative ought would most naturally predict different results: What Jaina ought to decide in this case, according to Gibbard's account of the deliberative ought, is a function of the non-normative information available to Jaina and the truth about PERMISSIBLE. Whether Jaina has enough credit to buy the meat is a non-normative matter. So, what Jaina ought to decide to do is dependent only on what is accessible to her about her credit. Jaina ought then to ask herself whether she believes that she has enough credit and whether any further information is accessible to her about it. A proper answer to that question fixes all of the non-normative information in light of which the ought that is being considered by Jaina applies. On the other hand, whether eating meat is permissible is a normative matter. So, what Jaina ought to decide depends not on Jaina's beliefs about that matter but on whether it is actually permissible.

Given this, Gibbard's account would seem to predict that the accessibility-based excuses that might apply to Jaina would differ based on whether the information being accessed is normative or not. CREDIT EXCUSE couldn't apply to Jaina's decision, Gibbard's account would predict, because the truth about CREDIT is irrelevant to whether Jaina decided properly. It's only what is accessible to Jaina about her credit that's relevant. The opposite is true for PERMISSIBLE though. According to Gibbard, the ought about which Jaina deliberates applies in light of the truth about PERMISSIBLE. So, if we can excuse poor judgment on the basis of facts not being accessible, we might be able to excuse Jaina by PERMISSIBLE EXCUSE. So Gibbard's account predicts the wrong result here. The intuition was that CREDIT and PERMISSIBILITY were on par in that Jaina's decision depends on the truth about one iff it depends on the truth of the other. Gibbard's account of the deliberative ought seems to suggest that when we evaluate that decision, we can rebuke her for failing to access one of the claims but not the other, even if she has equally good access to both.

Before moving on to the third worry for the asymmetry that Gibbard posits, it's worth considering a general explanatory schema that might be open to a defender of Gibbard here: One might think that we can explain the asymmetry Gibbard's account by appeal to a difference in how we learn the different types of information. Much of the non-normative information we learn, like the colors and sizes of everyday objects, is learned via our senses. In general, if we don't get the right sensory evidence, we wouldn't have learned that information. Evidence for the normative, on the other hand, doesn't seem to be something that we can "reach out and touch" in the same way. This asymmetry in our ability to access the information could ground an asymmetry in the information in light of which the deliberative ought applies.[24]

This kind of move can't account for the asymmetry, though. To show this, it's sufficient to show that at least some non-normative information is as accessible as some normative information. If that's the case, how accessible information is doesn't divide the normative and non-normative as the potential defender of Gibbard would require. By the objector's reasoning, if accessibility is the key to the asymmetry, the deliberative ought should either apply in light of the accessible non-normative information independently of the agent's beliefs about it or apply in light of the agent's beliefs about the normative.

To see that the normative can't be divided from the non-normative in terms of accessibility, consider Harman's (1977, p. 4) example where you "you round a corner and see a group of young hoodlums pour gasoline on a cat and ignite it." I take the point of this example to be that normative information, at least sometimes, is ascertained just like non-normative information; we "can *see* that it is wrong." Now suppose that the person rounding the corner were our deliberating agent. In terms of accessibility, the normative and non-normative information in this case seem to be on par.

Moreover, we often acquire non-normative information in very accessible ways, e.g. *a priori* reasoning to simple mathematical truths. In such cases, it would also seem that the non-normative information is at least as easily knowable as the answers to hard normative questions, like the permissibility of late-term abortions. So, a strict line between the normative and the non-normative cannot be drawn on the basis of accessibility to the agent. So, when it comes to figuring out which information the deliberative ought applies in light of, we can't appeal to just accessibility. If accessibility were to determine the divide between which information must be accessible to the agent and which need not be, then Gibbard's division between the normative and the non-normative would not result.

Some views of the normative take all distinctively normative facts to be accessible to all agents or at least accessible to ideally rational agents. The rational agreement requiring views, like Gibbard's (2003) view and ideal observer theories, are like this. These accounts would claim that the ideally rational version of Harman's subject does have access to a conditional like *that if there is a cat being*

[24] For example, you might think that something like this could explain the worry about Jaina's decision. We can't excuse Jaina's decision for her failure to get the right information about CREDIT since we can't hold Jaina to the standard of only using the truth about CREDIT in her decisions. We can't be blamed for not having non-normative information, in general, (the objector claims) because we could not have, on the basis of *our* evidence, ascertained it.

burned by young people, that is bad. So, this kind of view skirts around the worry posed here since they will claim that the normative should be at least as accessible to the subject as any non-normative information. I don't think this kind of view adequately captures what is going on in Harman's example, but after considering the third worry for Gibbard's asymmetry, I will show that the rational agreement requiring views are subject to a much more worrisome objection if they adopt Gibbard's proposed solution to the puzzle of correct belief.

The third and final point against rational disagreement permitting views that adopt GIBBARD'S ANALYSIS is more general than the previous two. The worry can be summarized like this: We'd expect that the difference between the subjective ought and the objective ought would be explained in terms of the difference between the subjective and the objective. Gibbard's view suggests that to properly characterize the subjective ought in contrast to the objective ought, the distinction between the normative and non-normative must also be drawn.[25] But, the contrast between the general notions of objectivity and subjectivity does not privilege the normative/non-normative distinction, I'll claim. So, that the subjective ought (in contrast to the objective one) importantly relies on the normative/non-normative distinction is, at best, surprising.

Ideally, to make this argument, I would offer a complete account of objectivity and subjectivity, but that is beyond the scope of the discussion here. Instead, I'll focus on the distinction between objective and subjective situations. Doing so should be sufficient to make clear that the subjective and objective oughts being different in terms of how they treat the normative and non-normative information is not plausibly due to what distinguishes them, namely the objective/subjective distinction.

What seems to be essential to the notion of subjectivity is the agent's ability to access what is subjective. That is, something is subjective when it's importantly tied to how things appear from the subject's perspective. To capture this idea, we can say that that some centered content is part of the *subjective situation* of an agent if it is internal, in the sense that the agent can tell that she is in the state, or it accessible from the inside. Some examples: being happy is part of the child's subjective situation since the child can tell that she's happy, i.e. she can "see" her

[25] The idea is that since the objective and subjective oughts treat the normative and non-normative differently, any way of distinguishing the objective from the subjective must account for this. Of course, an adequate picture of the subjective and objective might not distinguish the normative from the non-normative in those terms. Such an account may cut the terrain in a way that also splits the normative from non-normative divide without describing it as such.

happiness. Being (appeared to as if one is) in a red room is part of an agent's subjective situation if the agent has the experience as of being in a red room. Note though that this is not meant to be an intellectual account in the sense that some content is only internal if the agent *believes* that she's in it. Certainly one can be happy without having the concept of happiness or having any beliefs about being happy.

On the flip side, I'll say that some centered content is in the *objective situation* of an agent if it is true of the agent, regardless of whether it is accessible to the agent. That it's sunny and cool today in Ann Arbor is part your objective situation, even though you (probably) don't have access to that fact. Note that on this formulation, something could be part of both the objective and subjective situations of an agent.

According to Gibbard's view, when we consider the information that the *objective* ought applies in light of, there is no difference between the normative and the non-normative – the objective ought applies in light of the all that's so, regardless of what the agent can tell. But when we move to the *subjective* ought, the difference between the normative and the non-normative becomes relevant. Why would the normative/non-normative distinction become relevant when we switch from objective to subjective oughts? We'd expect that the distinction would enter only because it plays some kind of role in characterizing the subjective in contrast to the objective. I will argue that it doesn't play any such role. Because of that, we should be surprised if the distinction is actually is relevant.

Consider how the distinction between the normative and non-normative could sneak into the semantics of the 'ought's with the introduction of the subjective perspective. If the distinction comes with the subjectivity, then it must be because either the distinction plays a role in the characterization of subjective situations, i.e. in the distinction between the objective and the subjective, or the distinction plays a role in the subjective situations themselves, i.e. in the content of a subjective situation. But the normative/non-normative distinction doesn't play any special role in either of those, I'll now claim.

First, the normative/non-normative distinction doesn't appear to play an important role in the characterization of the subjective because the subjective can be characterized roughly in terms of what is accessible to the agent. What is accessible to the agent, in turn, is given by the true epistemology (of both the normative and the non-normative). But, as I suggested above, any plausible complete epistemology won't display a telling difference in accessibility between the normative and the non-normative. So, the normative/non-normative distinction doesn't play an important role in the difference between subjective and objective situations.

Next, the normative/non-normative distinction also doesn't play an important role in distinguishing the content of subjective situations. Of course, my claim isn't that the normative/non-normative distinction doesn't play an important role in *any* subjective situations. It does, for example, in the subjective situation of a metaethist who studies the distinction. But, for many, the distinction between the normative and non-normative is no more important to their subjective situation than the distinction between red and non-red objects. Certainly, our subjective experience does not, for example, automatically classify content as either normative or non-normative in a way that's peculiar to normativity (in contrast to other properties, such as redness). To see this, reflect on what it would feel like to be in Harman's cat example (given above, from Harman (1977, p. 4)). There, for example, we're receiving both normative and non-normative information (e.g. the color of the hoodlooms' clothes, their approximate height, the wrongness of the act, the blameworthiness of those involved), but the various bits of information seem to be on par in almost every way. All of the facts just appear to us in our subjective experience. Experience is not any more broken down into the normative and the non-normative than it is broken down into the red and not-red or the economic and non-economic. In fact, many even gloss over the distinction in thought and casual conversation, perhaps moving from a claim that something will increase happiness to those involved to the claim that they ought to do it, without recognizing that that is not merely an issue of the words involved.

So it appears then that the distinction between the normative and the non-normative is not drawn by subjectivity, either as part of subjective experiences or as part of the notion of subjectivity itself. Given that, it's hard to see why we should think that contrasting the objective and the subjective would draw such a line either. So, if the difference between the objective and subjective ought really is a difference between objectivity and subjectivity, that there would be an asymmetry in the subjective ought regarding normativity would be an unexpected result. This is the third reason to be suspicious of there being the asymmetry between normative and non-normative information that is required by rational disagreement permitting views that adopt GIBBARD'S ANALYSIS. It seems trivial that what *S* subjectively ought to do is what one ought to do in light of what's accessible in *S*'s subjective situation. But, since subjective situations are not asymmetric between the normative and the non-normative, subjective oughts are not either.

The three considerations I present show that rational disagreement permitting views ought to reject that the subjective ought is asymmetric with regard to which kinds of information it applies in light of. Each of the three worries I present presuppose that it's possible for there to be some normative truths that even an ideally rational agent might not have access to. As I mentioned above, rational agreement requiring views avoid these problems by denying that the normative is never really inaccessible to the ideally rational agent. These views face a different problem when trying to adopt Gibbard's solution to the puzzle of correct belief. I will explain that problem below. First though, I must slightly refine the arguments of this section, as they rely on a faulty assumption that the subjective ought is the deliberative ought.

2.4.4 The Deliberative Ought

When we deliberate about a decision, we're trying to decide what we *ought* to do. What we're looking for in deliberation then is an ought, but what kind of ought is it? I'll suggest here that the role of the deliberative ought is more complicated than it seems at first. Both the objective and my subjective ought can play the deliberative ought role in different contexts.

When I, in the first person, ask myself "What ought *I* do?", I'm deliberating about what to do. I'll call the ought that factors in this kind of deliberative thought the *first-personal deliberative ought*. When we're wondering about the first-personal deliberative ought, it seems that we want to know what we really ought to do in light of all that's so, not just what we ought to do given our limited knowledge of the world, I'll claim. To see this, suppose I am deciding what to do in a situation were I take the question of whether *P* to be highly relevant to what I ought to do, like Jaina does with CREDIT above. Were it the subjective ought that factors in our first-personal deliberation, we would expect that learning that I believe *P* is enough to close off consideration about the matter, since it's part of the very idea the subjective ought that it applies in light of the information available to the subject. But, when *P* is important for the decision, just learning that I believe *P* doesn't close off my wondering about *P*.[26] Were an omniscient advisor to approach me with the information that I believe *P*, I would still ask her: "Sure, I believe *P*. But is *P true*?" In other words, when we're asking ourselves "What ought I do?", what we believe about a fact does not screen off the fact itself. So when we're deciding what to do, the 'ought' in play is synonymous with the objective ought – what we want to know

[26] That is, learning that I believe *P* is insufficient for employing *P* in the decision-making process.

is what we ought to do in light of all the facts, normative and non-normative. It's the objective ought that plays the first-personal deliberative ought role.[27]

You might object that I've ignored the major motivation for thinking that the subjective, rather than the objective, ought plays the first-personal deliberative ought role. In the Three Envelope Case (or My FOUR ENVELOPE PROBLEM above), it's clear that one objectively ought to choose one of the two envelopes of unknown contents. So, if I'm right that we're deliberating about the objective ought, we should be able to rule out taking anything other than one of the envelopes of unknown content. Of course, though, were any of us actually in that scenario, we'd conclude that we shouldn't take either of those. Here, it looks like we coherently conclude both that we objectively ought to take one of the envelopes of unknown content but also that is not what we ought, in the first-personal deliberative sense, to do. So, it looks like I am mistaken in taking the first-personal deliberative ought to be the objective ought. That is, my argument seems to ignore the fact that we can only act in light of the information that we have, the objector claims.

In response, first notice that this problem is structurally analogous to the problem considered by Kolodny and MacFarlane (2010). While I will not endorse their semantics for 'ought' here, I will point to their work as an indication that this problem goes much deeper than my suggestion that the objective ought plays the first-personal deliberative ought role.

This second point will hopefully be more illuminating: notice that the claim I'm making here is that the 'ought' that plays the first-personal deliberative ought role is the objective ought. That is, the 'ought' in "What ought I do?" is synonymous with the ought that applies in light of all the facts. A separate (and false) claim would be that when we conclude thinking about the question of what to do, concluding "I ought to ϕ", the 'ought' in the conclusion is always the objective ought. Let's call this second ought, the one that occurs in the conclusions of our deliberations about what to do, *the first-personal decisive ought*. At least sometimes, like in the case above and in the case from Kolodny and MacFarlane (2010), we answer a question about what we objectively ought to do with an answer about what we subjectively ought to do. That is, the ought that plays the first-personal deliberative role is not always

[27] In this paragraph, I don't take myself to be disagreeing with Gibbard. Gibbard seems to make a similar point about the objective ought in (2005, p. 343): "It would be nice, to be sure, to know what objectively you ought to do. If indeed you could check with an omniscient advisor, that's what you would ask: If, say, you wonder whether to take an umbrella, who better to ask than someone with detailed foresight of whether it will be raining at moments you'll want to be outside?" Here I take myself to be clarifying the role that the objective ought plays in deliberation.

that same ought that plays the first-personal decisive role. Let me explain with an example:

Suppose Naomi wants to know whether it will rain today. To find out, she asks Emily, "Will it rain today?" Both Naomi and Emily know that it will either rain or it won't. But, none-the-less, Emily can felicitously answer "It's 70% likely to rain." Given what Naomi and Emily both know about the rain, this answer seems to be ruled out: it will either rain or it won't. Emily's answer is proper in this case though because it gives Emily's *estimate* of the probability of rain given the information available to her.[28] One might think that, strictly speaking, it doesn't answer Naomi's question because the answer is incomplete,[29] but we often interact like this. Likewise with the envelopes: when I ask myself which envelope to take, I want to know which one I objectively ought to take. Failing finding a proper answer to that, I'll take the best estimate of what I objectively ought to do, i.e. what I ought to do in light of the information available to me.[30]

So, the case does not generate a problem for my claim that the objective ought plays the first-personal deliberative ought role, as long as we realize that we don't always answer that question with a claim containing the objective ought. The first-personal deliberative ought is not the first-personal decisive ought.[31]

So we see that at least some of our conclusions about what to do contain the subjective ought in the first-personal decisive ought role, but we also have reason to think that *all* of our conclusions about what to do are like that. Objective oughts are true in light of all the facts, some potentially unknown to the agent to whom the ought applies. But, when we conclude a deliberation about what to do, the conclusion is directly motivating. Facts that I don't have access to can't motivate me. So, when I conclude a deliberation about what to do, the 'ought' of my conclusion must be one that applies in light of the information available to me.[32]

[28] I do not intend to be merely appealing to conversational norms, such as those explored by Grice (1975), to justify the properness of this response. By giving this answer, Emily is doing more than sharing all of the information she has on the topic; she's also potentially providing evidence for one of the binary answers to the original question. So, in this way, the answer is proper, but possibly *incomplete*.

[29] Here I'm employing a rough notion of questions and answers where posing a question sets up a partition of logical space. A complete answer picks out a cells of the partition, the one that the actual world is in according to the answerer.

[30] Instances of this kind of case are littered around the partial belief epistemology literature. For instance, we (subjectively) ought to have a credence of one-half in the fair coin having landed heads even though we know that it either did or didn't.

[31] This general line might also be able to be developed into a yet unconsidered response to the problem posed by Kolodny and MacFarlane (2010).

[32] We can also see this by using the method of screening off appealed to above. If I conclude that

We see then that the first-personal decisive ought must apply in light of the information available to the agent. Since facts that I don't have access to can't motivate me, regardless of whether those facts are normative or not, there doesn't appear to be an asymmetry between the non-normative and normative here. As such, the ought that plays the first-personal decisive role is *my* subjective ought. When we make a conclusion about what we ought to do, the 'ought' in that conclusion is the ought that applies in light of what the agent takes to be the case both non-normatively and normatively.

To recap, we see that there is no single ought that factors in deliberations about what to do. When we ask ourselves what to do, we're wondering about what we objectively ought to do. In the case of belief where we wonder whether P, we can conclude that probably P or probably not-P on the basis of the information available to us. Likewise for deciding what to do, we answer the question of what we objectively ought to do with a subjective ought, one that applies in light of the information available to us.

In the previous section, I gave three reasons for thinking that the subjective ought, conceived of as a univocal ought about which we deliberate, does not respect an asymmetry between the non-normative and the normative. Now that we see that there is no unique ought that plays in our deliberation, the arguments above appear to rely on a false premise, and the question of the asymmetry may be resurrected. But, a careful look at the three arguments given will show that the first two only rely on the ought playing the decisive deliberative role, whereas the last only depends on it being subjective rather than objective. Both of those assumptions are supported by the arguments in this section. As such, the three points above collectively serve to show that neither the subjective ought nor the first-personal decisive ought (which, I claim, is the same as the subjective ought) respect the asymmetry. Since the only other ought that factors in deliberation is the objective one, I conclude that no ought that serves in deliberation treats the normative differently from the non-normative. So, advocates of views that allow for fully rational agents to disagree about distinctly normative issues ought to reject Gibbard's solution to the puzzle of correct belief; Gibbard's solution commits them to an awkward asymmetry between

I ought to ϕ, the 'ought' of my conclusion better be one where my subjective situation screens off the facts, i.e. my beliefs about my situation give me as much information as I could need to decide whether the 'ought' applies. Else, I couldn't always be motivated by the 'ought' judgment. Gibbard seems to agree on this point, and I take this kind of reasoning to be his motivation for thinking that there is a unique deliberative ought that is the subjective ought (Gibbard 2005, p. 343). I expand on this discussion below when I claim that judgments about my subjective ought are the motivating ones.

how normative and non-normative information is handled by the subjective ought.

2.4.5 Treating the Normative as Rationally Accessible

The points in the previous section show against any account of the subjective ought that adopts Gibbard's solution to the correct belief puzzle and allows that some normative facts may be inaccessible to an ideally rational agent. My arguments rely on it being possible that there is some normative information that is not accessible to the rational agent. Some accounts of the epistemology of the normative will avoid the above problems by denying that the normative considerations in light of which the subjective ought applies can ever really be inaccessible to an agent, at least in the limit of rationality. So in any subjective situation, an ideally rational agent would have access to the normative facts, and they can factor in the subjective oughts, the accounts claim.[33]

These accounts – the rational agreement requiring accounts – are subject to a different worry though. I will argue here that if a rational agreement requiring view adopts Gibbard's solution to the puzzle of correct belief, it is forced to make the subjective ought so inaccessible to some agents that it can hardly be considered *subjective*. The problem is generated because what is accessible to the ideally rational version of an agent is often not accessible, even in a weak or limited sense, to the actual agent. Subjective oughts, I claim, cannot be this inaccessible.

2.4.5.1 An Accessibility Constraint on Subjective Oughts

In the literature on subjective oughts, the subjective ought is meant to play a number of different roles. Holly M. Smith's "Subjective Rightness" (2010, p. 72) catalogues and discusses a number of these roles.[34] The roles that the subjective ought is meant to play include a few already mentioned here (either explicitly and implicitly): The subjective ought is used to (1) explain how an action can be right in some sense, such as the three-envelope problem, even though the action is also wrong, in some sense (Smith calls this "Normative Adequacy"); (2) provide a sense of 'ought' that's connected to blameworthiness in a way that objective oughts are not (which Smith calls "Relation to Blameworthiness"); and (3) be an 'ought' that guides and

[33] See note 15 above about how Gibbard's view fits into this camp.

[34] Smith concerns herself with the analogous notion of subjective *rightness*, rather than the subjective *ought*, and particularly considers that kind of rightness in the moral realm. The same ideas apply in the broader normative realm and concerning the subjective ought.

motivates agents to act. I'll call the last role "the guidance and motivational role."[35]

To generate the disconnection worry that I'll leverage against the rational agreement requiring views below, I'll need to say more about the final role for the subjective ought, that of being the ought that can guide and motivate agents. The guidance and motivational role consists of both a guidance aspect and a motivational aspect. For an ought to satisfy the guidance aspect, it must be that the agent can be guided by the content of the subjective ought claim. To borrow Smith's example, if an agent accepts that she subjectively ought to stop at red lights, then if the agent wanted to act in accord with the principle, she'd be able to derive a way to act in accord with it.[36] Smith (2010, p. 73) notes that this kind of guidance criterion is ambiguous between an internal and external characterization: on the external characterization, subjective ought claims guide if they provide the agent a way of acting that makes his actions actually accord with the principle. On the external understanding, the subject of the subjective ought claim about red lights is guided by it if she can stop at red lights in virtue of accepting the principle. In general, normative principles are not like this though – we can't conform our action to a principle in virtue of accepting that principle. Instead, we should think of the subjective ought as internally guiding in that the agent can find some prescription for action that at least appears to help the agent conform his behavior to the principle, regardless of whether it succeeds. Smith (2010, p. 73) argues that we can at least expect the subjective ought to be internally guiding:

> [It] seems realistic to insist that principles of subjective rightness – which, after all, are designed to guide agents in making decisions when they are mistaken or uncertain about what the governing principle of objective rightness requires of them – should at least be capable of being used as internal decision guides. An agent who cannot find any way to translate his moral values into his choice of what to do is an agent who cannot find a way to govern his decision by the considerations he deems most relevant. His decision does not express his moral values, and so in an important way undermines his autonomy. Thus, we want principles of subjective rightness to be capable of being used as internal guides to action, even if they cannot successfully be used as external guides to action.

[35] The final role I assign to the subjective ought doesn't map on nicely to any of Smith's. It includes her "Guidance Adequacy," but it is strictly stronger in that it includes a motivational aspect.

[36] This is very close to Smith's characterization of the similar role on page 73.

Following Smith, I'll take it that subjective oughts do meet this internal guidance criterion: If an agent subjectively ought to follow some principle, then the agent is capable to deriving some prescription for action from that principle (though following that prescription may not guarantee that the agent's actions conform to the principle). Gibbard seems to accept something at least this strong in requiring that the norms that we accept be couched in recognitional terms (Gibbard 2002).

The guidance aspect of the guidance and motivation role requires that the subjective oughts that apply to us can guide our action, at least in an internal way. The motivational aspect is supposed to capture what Gibbard claims about the subjective ought, that they "are the oughts that exert normative governance – the oughts we accept and whose acceptance is directly motivating" (2012, p. 80). The motivational aspect, unlike the guidance aspect, requires that if an agent accepts a subjective ought, the agent is motivated to conform to it. Gibbard points out that the subjective ought playing this motivational role is reason to think that it applies only in light of the information available to the agent, saying, "We must act, after all, in light of information we have" (2012, p. 80). This is, in part, how our acceptance of a subjectively ought claim can motivate us – it applies in light of the things we take to be true.

It is important to clarify this last claim, that subjective oughts motivate us in virtue of applying in light of what we take to be true. One might think this claim omits an important aspect of how how subjective ought claims work, that they often apply in virtue of our ignorance, as well, like in the envelope case above. A few modifications must be made then: First, if agents' attitudes come in degrees, then we can say that the ought claims apply in virtue of the agents' degreed attitudes, regardless of whether those constitute acceptance, rejection, or suspension of belief. Also, if we have non-belief attitudes that represent their content as true, the subjective ought should apply in light of the contents of these attitudes as well. Finally, even if agents' attitudes come in degrees, agents can fail to have even intermediately degreed attitudes in some cases – contrast, for example, having an intermediate credence in whether the coin will come up heads with having no credence at all about that. Subjective oughts may also apply in light of the having of an attitude or lack thereof. Together, we can think of these things, all of the agent's attitudes that represent their content as true to the agent, as the agent's picture of reality. This then is what the subjective ought applies in light of. The subjective ought can motivate because its prescriptions are based in how we take things to be, in this wide sense. Below, when I discuss 'what an agent takes to be true' or similar notions, I will be referring to the agent's picture of reality, in this sense.

With this idea in mind then, we can formulate the motivation aspect of the subjective ought role with two prongs: first, when we accept a subjective ought, we're motivated to accord our action with it, and second, acceptance of subjective oughts motivates us because they apply in light of what we take to be true. Following Smith and Gibbard, I'll assume that the subjective ought plays the guidance and motivation role. This results in the following accessibility condition on the subjective ought:

ACCESSIBILITY *S* subjectively ought to ϕ only if

(a) *S* is capable of being guided by that subjective ought – that is, *S* is capable to deriving some internal prescription for action from that *S* subjectively ought to ϕ,

(b) if *S* were to accept that *S* subjectively ought to ϕ, *S* would be motivated to ϕ because the ought applies in light of what *S* takes to be true.

Together, these two requirements put an accessibility requirement on subjective oughts. For it to be the case that an agent subjectively ought to do something, two conditions must obtain: that principle must be accessible to the agent in that the agent is able to try to follow it, and the agent would be motivated to try to follow it if she were to accept that she ought to and her motivation is in part explicable because the ought applies in light of what the agent takes to be true.

One might worry that the accessibility condition I've given here is still too strong. Consider this subjective ought, for example:

> Sandra, who is 5 years old, thinks that the chance of winning a dollar in the lotto is over .5, and she thinks that having a dollar is more than twice as good as having a quarter. So, she *ought* to choose the lotto ticket over the guaranteed quarter.

In this case, as described, it seems that Sandra subjectively ought to choose the ticket, even if Sandra is not in a position to have any notion of anything like the expected value of the ticket, perhaps because she's lacking the relevant concepts. The worry would then be that because Sandra lacks the relevant concepts, she's not in a position to be guided by that directive to buy the ticket.

This worry, as stated, is not a worry for condition of ACCESSIBILITY given above. All that ACCESSIBILITY requires is that Sandra be able to be guided by the directive to buy the ticket and be motivated by accepting it. This doesn't require that Sandra be able to reason herself into a position where she tell *why* she ought to buy the ticket. All it requires is that she able to deduce some action that, by her own lights, helps her conform her action to the directive. She need not have the concept EXPECTED UTILITY to do that.

There is a nearby worry though. Consider an agent who starts out with a incoherent collection of beliefs: Suppose Sandra, now an adult, is again offered the lotto ticket for a dollar payoff or the quarter. Suppose that she believes the ticket is .5 likely to payoff, that it's more than two times better to have a dollar than to have a quarter, and accepts the norms of standard rational decision theory. Suppose further that Sandra does some quick (and mistaken) math and concludes that she ought to take the quarter.

Even though Sandra believes that she subjectively ought to take the quarter, it seems that Sandra really subjectively ought not take the quarter. This is incompatible with ACCESSIBILITY because ACCESSIBILITY requires that were the agent to accept that she subjectively ought not take the quarter, then she'd be motivated to refrain from doing it because that ought applies in light of what she takes to be true. But, Sandra takes it to be true that she ought to take the quarter. So, in light of what Sandra takes to be true, Sandra ought to take the quarter. This is just an instance of the trivial logical schema appealed to in note 17 above, IN-LIGHT-OF REPETITION: For all P and Σ, if $O(P) \in \Sigma$, then in light of Σ, $O(P)$, where O is the 'ought' operator.[37] So, assuming that it can't be the case that she both ought and ought not take the quarter, the 'ought' that tells her to refrain from taking the quarter does not apply in light of what Sandra takes to be true. Strictly speaking then, that Sandra subjectively ought not take the quarter is precluded by the motivation aspect of ACCESSIBILITY as it stands.

We think though that Sandra subjectively ought to give up on her belief that she ought to take the quarter. Why is this? It's because taking the ticket is required by what Sandra is really committed to (namely, her beliefs about the likelihood of

[37] If Sandra is to be motivated by accepting that she subjectively ought not take the quarter because it applies in light of what she takes to be true, then Sandra would have to be committed to a contradictory collection of beliefs about what she ought to do (namely, take only the ticket and take only the quarter). I'm assuming it's possible to accept a contradictory collection of beliefs about what one ought to do. We often do this. For example, I think I should treat each of my students equally. This means that each of them should receive the same amount of time on the test. But Sunny has a disability that warrants that he receive extra time. I also think that students with disabilities ought to be accommodated. In doing so, I accept a contradictory collection of norms. That said, we typically have ways of revising what we accept in these cases (but that doesn't threaten my claim). Notice also that this is strictly weaker than the much more controversial claim that there are real ethical dilemmas. The latter claim looks much less plausible.

winning and the relative values of the possible outcomes). Her belief that she ought to take the quarter is the result of mistaken math, not her considered judgments. So, to make sense of cases where agents have mistaken views about what they subjectively ought to do or the more general case where agents are committed to norms that are in tension with each other, we should revise the accessibility requirement to WEAK ACCESSIBILITY:

WEAK ACCESSIBILITY *S* subjectively ought to ϕ only if

(a) *S* is capable of being guided by that subjective ought – that is, *S* is capable to deriving some internal prescription for action from that *S* subjectively ought to ϕ, and
(b) if *S* were to accept that *S* subjectively ought to ϕ, *S* would be motivated to ϕ because the ought applies in light of what would be *S*'s considered judgments about what is true.[38]

This is an accessibility constraint that any subjective ought must satisfy. Subjective oughts must be able to guide the agents to whom they apply, and they're able to motivate the agents who accept them because they apply in light of the considered judgments of the agent. I'll show in the next section that rational agreement requiring views that adopt Gibbard's proposed solution to the puzzle about correct belief cannot give an account of the subjective ought that satisfies WEAK ACCESSIBILITY.

2.4.5.2 The Disconnection Worry

I will show in this section that rational agreement requiring views that adopt Gibbard's solution to the puzzle of correct belief, which takes the subjective ought to apply in light of the true norms, must deny WEAK ACCESSIBILITY. Showing that these views must deny one of the conjuncts of WEAK ACCESSIBILITY is enough to prove the claim, but I'll show that the views must actually deny both of them.

First, consider my super honest cousin Jeb. Jeb thinks that it's impermissible to lie under any circumstances. Jeb accepts many of the other norms that we accept, including normal epistemic norms, but he isn't nearly as committed to those as

[38] To make sense of what would be *S*'s considered judgments about what ought to be true, we can employ the notion of a reflective equilibrium, but this isn't strictly speaking necessary. Also, you might think that this condition should be that the ought applies in light of what *should* be *S*'s considered judgments. This 'should' must be the subjective one, if we're to retain the subjective aspect of the accessibility constraint. That's not a problem for the view being presented, since it isn't intended to be a reduction or analysis of the subjective ought.

he is to the norm against lying. So, no amount of reasoning would make him change his mind about lying.[39] So by Jeb's lights, the best thing to do if a known killer is at his door seeking information about his friend is to not lie. Because Jeb feels to strongly about lying and not nearly as strongly about any other norm (or collection of norms), this is part of Jeb's considered judgments about what the correct norms are. Presumably though, if the killer is going to kill an innocent person, then *objectively speaking*, Jeb ought to lie to the killer.

Now apply GIBBARD'S ANALYSIS to this objective ought: that Jeb objectively ought to lie amounts to it being the case that were it that Jeb subjectively ought to accept all that's so, it would be the case that Jeb subjectively ought to lie. Notice that given ACCESSIBILITY, it can't be the case that Jeb actually subjectively ought to lie. That's because that subjective ought is incompatible with the motivation aspect of the subjective ought role, that if S were to accept that S subjectively ought to ϕ, S would be motivated to ϕ because the ought applies in light of what would be S's considered judgments about what is true. It's part of Jeb's considered judgments that he subjectively ought not lie. So in light of his considered judgments, he subjectively ought not lie. Were he to accept that he ought to lie, it would not motivate him because it applies in light of his considered judgments. In fact, it blatantly contradicts his considered judgments. By the subject's lights, it can't be the case that he ought to lie, so it's not the case that he subjectively ought to lie.

What then about the counterfactual where Jeb subjectively ought to accept all that's so? Recall that for Jeb to accept all that's so, in the sense GIBBARD'S ANALYSIS intends it, is for him to accept "a full and true factual description of [his] circumstances" (2012, p. 83). That is, Jeb accepts all that's so when he accepts all of the non-normative facts that are true of him and his situation.[40] If Jeb, our super honest fellow, were to do this, it hardly seems he'd give up on his commitment to honesty. Jeb, you'll recall, is committed to not lying more than he's attached to

[39] Really, it's not that he *wouldn't* change his mind. It's that changing his mind doesn't improve his views by his own lights. So, he wouldn't change his mind in a way that he would now approve of.

[40] One might think that the best way to interpret the antecedent of the conditional is as talking about the possibility were one subjectively ought to believe *all* that's so, including all of the normative facts. Gibbard explicitly rules this out. But also, it might make objective oughts too easy: Suppose that S subjectively ought to accept that she objectively ought to ϕ. That means that in light of what is available to her, she ought to think that she objectively ought to ϕ. But if what is available to her suggests *thinking* that she (objectively) ought to ϕ, it's easy to imagine that what is available to her would also suggest ϕ-ing. If so, then she subjectively ought to ϕ. So, it seems that if the antecedent of GIBBARD'S ANALYSIS is supposed to include *all* information including facts about what the subject objectively ought to do, then GIBBARD'S ANALYSIS would entail that many, if not all, objective oughts come for free.

anything else he believes. Surely, it would be a strange scenario, perhaps one in which Jeb subjectively ought to believe things that aren't suggested by his evidence. But, it doesn't look like a situation where Jeb would be required to change his deep-seated beliefs about lying. So, were he to accept all of the non-normative facts, it still seems that it'd be part of his considered judgments that he ought not lie. Then, by the above reasoning again, it won't be the case, even in this counterfactual scenario that Jeb subjectively ought to lie. We see then that if WEAK ACCESSIBILITY is true, GIBBARD'S ANALYSIS predicts the wrong results about Jeb. GIBBARD'S ANALYSIS is incompatible with the motivational aspect of the subjective ought role.

The example of Jeb relies on there being agents who are deeply committed to false normative claims. Jeb is *deeply committed* to not lying, in that his normative commitment is stable in the sense of Egan (2007). He is deeply committed when "no change that the believer would endorse as an improvement would lead them to abandon it" (2007, p. 212).

According to the rational agreement requiring views, ideally rational agents believe only true normative claims. Given this, advocates of rational agreement requiring accounts have two options: they must either think that we can be deeply committed to false normative claims or deny the possibility of agents like Jeb. Denying the possibility of agents who would have false normative commitments under any change that they would endorse as an improvement in their normative attitudes would be rash, and Gibbard doesn't do this.[41] It is at least possible that we may be deeply confused or committed in ways that simply don't permit us to overcome our normative mistakes merely by deliberation. So that leads the views under consideration to deny that we can reach ideal rationality by improvements we would deem to be improvements in our own attitudes.

Denying that we can reach ideal rationality just by changes in our attitudes that we would deem to be improvements may not seem like such a bad option. This leaves the view saying that the subjective ought really applies (and would be accessible to) the ideally rational agents, not us mere mortals. But, this move makes the subjective ought too detached from the agent.

In order to be guiding and motivating, subjective oughts must be applicable in light of what the agent takes to be the case.[42] But if an advocate of a rational agreement requiring view accepts GIBBARD'S ANALYSIS to solve the puzzle of

[41] See footnote 15 for a discussion of in what sense Gibbard accepts this.

[42] I take Gibbard to characterize the subjective ought like that as well. Gibbard says, "The basic normative precepts that ground a subjective ought are subjectively applicable – applicable in light of information the agent has" (2012, p. 76).

correct belief, she is forced to think that subjective oughts are really only subjectively applicable to the ideally rational versions of the agents to whom they apply, and the ideally rational versions of agents may be markedly inaccessible to the agent in that the idealized agent wouldn't even be an improvement of the agent by the agent's lights.

The marked inaccessibility of ideally rational agents on this view is particularly stark in the case where actual agents lack the concepts involved in the subjective ought claims. Take the case of Luke, who has spent his whole life in the detached village of Peaceville. Peaceville has never had any criminal activity, so that possibility rarely, if ever, crosses the residents' minds. Luke, in fact, doesn't even have the concept CRIMINAL, so he can't even conceive of the idea.[43] On Tuesday, Luke is crossing the street when he sees John breaking the window of the local jeweller and quickly putting all of the merchandise into a bag. Of course, John is criminally robbing the jeweller. Suppose that the concept CRIMINAL is a thick concept consisting of the non-normative quality of violating the law and the normative quality of breaking the law being wrong.[44] Luke does not see it as a criminal matter though, because Luke doesn't have the requisite concepts.

In the scenario under consideration, we can imagine that it is the case that Luke ought to accept a full non-normative picture of how things are, at least with regard to the robbery. Luke has the relevant perceptual justification for believing all of the non-normative facts about the robbery.

Suppose now that Luke finds himself as the prosecuting lawyer at the trial for John, still lacking the concept CRIMINAL. Objectively speaking, the best thing for Luke to do is to convince the judge that John has committed a criminal act. So, according to the solution under consideration, were it that Luke subjectively ought to accept a full non-normative picture of how things are, it would be the case that Luke subjectively ought to convince the judge that John has committed a criminal act. But here, it is the case that Luke subjectively ought to accept a full non-normative picture of how things are. Hence, according to the solution, Luke subjectively ought to convince the judge that John has committed a criminal act, even though he doesn't even have the concept CRIMINAL.

But without the concept CRIMINAL, it doesn't seem that Luke could even *try*

[43] I'm assuming here that concepts are the constituents of thoughts. If one lacks a concept of something, then she cannot think about it. I'm also assuming, against views like Fodor (1975), that we can lack concepts. If I am wrong that we can lack concepts, this example can be reformulated into a case where Luke fails to ever apply the concept.

[44] For discussion of this, see Elgin (2005, p. 343).

to convince the judge of this.[45] WEAK ACCESSIBILITY requires that subjective oughts be able to guide the agent, in that if S subjectively ought to ϕ only if S is capable to deriving some internal prescription for action from that S subjectively ought to ϕ. In this case, Luke cannot derive such a prescription for action. Luke cannot even think the thought that he ought to convince the judge that John is has committed a criminal act, so how could he possibly derive a prescription for action from it? This combination of views requires that subjective oughts violate WEAK ACCESSIBILITY, so subjective oughts on this picture are too disconnected from the subjective situation of the agents to whom they apply.[46]

One might object that my WEAK ACCESSIBILITY is too strong. In the case of Jeb, we see that WEAK ACCESSIBILITY entails that if an agent is deeply committed to it being the case that she subjectively ought to ϕ, then it can't be the case that she subjectively ought not ϕ. Roughly, this says that agents can't be deeply mistaken about what they subjectively ought to do. If that's right, then one might object that WEAK ACCESSIBILITY entails that subjective ought claims aren't normative. When I say that Jeb subjectively ought not lie, it seems that I'm just reporting something about Jeb's mental state, namely his deep commitment to not lying; I'm not making

[45] One might think that the actions prescribed by ought claims can always be formulated in purely non-normative terms. This might avoid my criticism as the antecedent of the conditional of the solution would require the the agent have the relevant conceptual scheme. For example, in this case, one might say that what Luke really subjectively ought to do is make the sounds and arm movements required to get the judge to accept the sentence in his own language that means that John is a criminal. Plausibly, John could conceptualize this if he subjectively ought to accept a full non-normative account of how things are since that account will include facts about the judge's mental states and how they affect whether John is punished by the state. I reject this move though, and here's why: Even though making the sounds and arm movements required to get the judge to accept some sentence actually corresponds to convincing the judge that John is a criminal, these two things have different modal profiles. The action that John ought to do is convince the judge that John is a criminal, even in possible worlds where this doesn't line up with the non-normative characterization.

[46] The subjective ought is so detached from subjective situations on these accounts, that it can advise agents to ignore their own considered judgments about facts in light of their evidence. Take a case analogous to Jeb's above: If Raul evaluates his evidence as indicating that P is mostly likely, even if he would maintain this as a considered judgment, the subjective ought of these accounts would permit that Raul subjectively ought to judge that P is false. This would happen when Raul is deeply committed to non-optimal epistemic principles, such as sharing information with as many colleagues as possible. (See Zollman (2007), where he shows some surprising results about what kinds of communication networks best promote our epistemic goals.) It seems reasonable to think that when Raul is mistaken about how to respond to his evidence, then he *objectively* ought to judge it differently than he would think he should. But if Raul is committed to thinking that P is supported by his evidence because he is deeply committed to sharing information with colleagues, it's difficult to imagine that Raul subjectively ought not think that P. This certainly flies in the face of the spirit of subjectivist partial belief views, such as de Finetti (1937) and Jeffrey (1983). Gibbard also appears to think that questions about how to respond to one's evidence are intended to be handled by the subjective ought and highly dependent on the subject's views in this way (2012, p. 80-82).

a normative claim, the objector claims.[47] Given WEAK ACCESSIBILITY, the objection is that claims about what an agent subjectively ought to do merely report what the agent is deeply committed to, not what the agent ought, in any normative sense, to do.

In one respect, the objector is right: in some cases, the subjective ought is restricted by the non-normative, such as when the agent is deeply committed to a claim about what she subjectively ought to do. In response to the worry though, notice that there being some non-normative way of picking out some instances of the property that 'subjective ought' picks out isn't threatening to the view that the subjective ought is normative. What would be more threatening is if WEAK ACCESSIBILITY entailed that the subjective ought *means* something about the agent's mental states. Of course, this is false, for two reasons. First, WEAK ACCESSIBILITY doesn't make any claims about what the subjective ought means; it merely gives a necessary condition for its application. Second, even if WEAK ACCESSIBILITY were a claim about meaning, it would not be even close to complete. Most of the time, we have either no views or at most partial views about what we subjectively ought to do. Even then, we're typically more committed to other norms (cast in non-subjective terms) that can override our subjective ought beliefs when we fully consider our commitments. If we've promised to pick up the kids from school on time and finish the project at work on time, when we come to realize that there's not enough time to do both, we're in general more deeply committed to our non-normative beliefs than we are to views about what we ought to do.[48] Cases where an agent's considered commitments deliver judgments about what to do are rare. So, in general, when the subjective ought is used to prescribe what an agent ought to do, its being normative is not threatened by the requirements of WEAK

[47] A possible line of response to this worry (which I accept) is inspired by Shah (2003) and Evans and Shah (2010). Shah argues that facts about what an agent believes are normative facts. Evans and Shah (2010) consider the possibility that all mental states are similarly normative. If that's right, then making a claim about an agent's mental state is making a normative judgment. So, the purportedly non-normative subjective ought facts are genuinely normative even when they are fixed by the facts about the agent's mental states because those facts are normative too. Gibbard (2012) rejects the second solution because he thinks it gives an inaccurate account of belief. Gibbard also takes his explanatory project in the chapter to hinge on explaining belief in terms of sentence acceptance, a non-normative notion. But this requirement of the project assumes that the objective ought can be explained in terms of the subjective ought. The arguments above show that there is no acceptable view of the normative compatible with Gibbard's way of explaining the objective ought in terms of the subjective ought. So, the above considerations cast doubt on the proposed explanation.

[48] I'm assuming here that we're also committed to some norms about how to resolve conflicts or seek advice. This isn't strictly speaking necessary though. Any case where the subject's views don't deliver a considered judgments about what the subject subjectively ought to do will be sufficient.

ACCESSIBILITY.

Rational agreement requiring accounts that employ Gibbard's solution to the puzzle of correct belief must make the subjective ought so markedly disconnected from the subjective situation of the agent that it can't motivate or guide the agent, so it can't play the role we want the subjective ought to play. The subjective ought, on these views, violates both conjuncts of WEAK ACCESSIBILITY as shown with the cases of Jeb and Luke. Since any account of normative epistemology must either be rational disagreement permitting or rational agreement requiring, we see that no way of adopting Gibbard's solution is acceptable. The first class is left without an explanation of the asymmetry it must posit in the subjective ought, and what the second class calls 'the subjective ought' can hardly be considered subjective. Since no epistemology of the normative can nicely account for the subjective ought and adopt Gibbard's solution to the puzzle of correct belief, we ought to reject Gibbard's solution.

2.5 Solving the Puzzle

Here again is the puzzle about correct belief that motivated GIBBARD'S ANALYSIS:

(1) Snow is white.
(2) Sophia believes that snow is white.
∴ (3) Sophia's belief that snow is white is correct.

The autonomy of ethics is supposed to block valid inferences of a normative claims from non-normative premises. The puzzle is to show explain how this argument can be valid without denying the autonomy of ethics. A benefit of Gibbard's approach to the puzzle was that it aimed to save the appearances: it tried to maintain both the validity of the argument and the normativity of the conclusion while squaring it with the autonomy of ethics. Unfortunately, we ought to reject Gibbard's solution for the reasons given above.

In place of Gibbard's solution, I propose a different trivial conclusion response. My explanation the puzzle employs the insight of Shah (2003), that the concept BELIEF is normative. On this view, it's a conceptual truth that a belief that P is correct if P. Using this conceptual truth, the puzzle can be explained easily. The autonomy of ethics is not violated because the argument contains a normative premise, the one that ascribes belief to Sophia.[49]

In the Chapter III, I argue for NORMATIVE TRUTH REGULATION:

NORMATIVE TRUTH REGULATION It's a conceptual truth about belief that if some attitude is a belief that P, one ought, other things being equal, have that attitude if and only if P.

If we take 'correct' in this context as normative and think of the correct belief as the belief that one ought to have, then my claim nicely folds into Shah's.[50] So, with the argument I provide in the next chapter, we have the resources needed to explain the puzzle. It's part of the very idea of belief that beliefs ought to be true. So, when we ascribe Sophia the belief that snow is white, we're making a normative claim. The autonomy of ethics is not violated because the argument contains a normative premise.

[49] Recall that my formulation of the puzzle differs slightly from Gibbard's. Gibbard's version does not contain a normative premise, but the conclusion is trivial given the premises. These kinds of arguments don't threaten the autonomy of ethics in the same way that it isn't threatened by Prior's (1960a) examples. Given the premise that snow is white, then it falls out as trivial that Sophia's belief that snow is white is correct due to the meaning of 'belief'.

[50] Treating 'correct' in this way is done by Shah and Gibbard as well. The idea is that, following Ewing (1939), we can say that the *correct* belief is the belief we ought to have, for some sense of 'ought.'

CHAPTER III

Why Believe the Truth

ABSTRACT

Clifford (1877) claims that we have a *moral* obligation to accord our beliefs with our evidence. What to believe, at least about mundane topics, doesn't seem like a moral issue, though. What we ought to believe also doesn't seem to be determined by our desires. But, what else could ground facts about what we ought to believe? I offer a new account of the truth norm, which can play this role. At the heart of the account is the idea that the concept BELIEF is normative. It's a conceptual truth, I claim, that other things being equal, beliefs ought to be true. In support of this claim, I show how it can give a compelling explanation of an aspect of knowledge ascriptions that is otherwise difficult to explain. I then claim that being an agent requires being subject to this norm of belief. This results in a non-moral, distinctly doxastic, account of why we ought to believe the truth. My conclusion is that asking why we ought to believe the truth is like asking why a bachelor must be unmarried: the answer is contained in the ideas that make up the question.

William K. Clifford starts out "The Ethics of Belief" (1877) with the example of the shipowner:

> A shipowner was about to send to sea an emigrant-ship. He knew that she was old, and not overwell built at the first; that she had seen many seas and climes, and often had needed repairs. Doubts had been suggested to him that possibly she was not seaworthy. These doubts preyed upon his mind, and made him unhappy; he thought that perhaps

> he ought to have her thoroughly overhauled and refitted, even though this should put him at great expense. Before the ship sailed, however, he succeeded in overcoming these melancholy reflections. He said to himself that she had gone safely through so many voyages and weathered so many storms that it was idle to suppose she would not come safely home from this trip also. He would put his trust in Providence, which could hardly fail to protect all these unhappy families that were leaving their fatherland to seek for better times elsewhere. He would dismiss from his mind all ungenerous suspicions about the honesty of builders and contractors. In such ways he acquired a sincere and comfortable conviction that his vessel was thoroughly safe and seaworthy; he watched her departure with a light heart, and benevolent wishes for the success of the exiles in their strange new home that was to be; and he got his insurance-money when she went down in mid-ocean and told no tales.

Clifford uses this kind of example to defend the principle that "it is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence." Importantly, Clifford's principle is supposed to be a *moral* one. We're morally obliged to believe only with sufficient evidence, according to Clifford, since "it is not possible so to sever the belief from the action it suggests as to condemn the one without condemning the other." Clifford also takes his dictum to hold for all beliefs, since they all have "some influence upon the actions of him who holds it."

The norm that Clifford endorses is an *epistemic norm* in that it gives agents accessible instructions about what to believe. In contrast, William James commands us to "Believe truth! Shun error!" (1979, p. 24). Those commands set the *doxastic standard* – a standard that agents may not be in a position to immediately satisfy. Many authors take there to be a natural connection between James's doxastic standard and epistemic norms like Clifford's. The doxastic standard sets the goal for belief while epistemic norms tell us how to reach it.[1] Following that line, we'd expect that if we have a moral obligation to believe only with sufficient evidence, it's because we have a moral obligation to believe the truth.

In the example above, the shipowner not proportioning his beliefs to his evidence does have morally grave consequences, but it's odd to think that Clifford's norm is a moral one. To borrow an example from Tom Kelly (2003, p. 262), if before you've seen the newly-released film, your colleague blurts out that the main character dies

[1] See Owens (2003, p. 283) for a discussion of the relationship between the two types of norms.

at the end, you should believe her. But, it's hard to think that you have a particularly *moral* duty to believe her, as Clifford would predict. If you ignore her testimony, you've done something wrong, but you hardly seem unethical.[2]

One might try to replace Clifford's appeal to the moral with an appeal to the pragmatic: perhaps we ought to accord our belief to the evidence because doing so best promotes our goals. But this kind of account won't work either. Sometimes we have beliefs about things that are completely independent of our goals, like my belief about what my phone number was 15 years ago. We are also subject to epistemic norms even when they conflict with our goals. Kelly's (2003, p. 262) movie example shows this too: even if you have the goal of not finding out the ending to the movie so you can enjoy it when you see it, you must apportion your belief to the evidence after your colleague blurts out the ending. So a pragmatic account of these norms, like Clifford's moral story, doesn't look sufficiently general either.

My goal here is to offer an account of why we ought to believe the truth that doesn't treat it as a moral or pragmatic issue. The account proceeds in two steps: First, I endorse a position advanced by Shah (2003) that the concept of belief is normative. I offer new support for the normativity claim by showing that it offers a compelling explanation of an aspect of knowledge attributions that is otherwise difficult to account for. Second, I show that being an agent requires being subject to the norms of belief. Since the norm "Believe the truth" is meant to apply to all and only agents, that gives us a non-moral, distinctly doxastic, account of why we ought to believe the truth. My conclusion is that asking why we ought to believe the truth is like asking why a bachelor must be unmarried: the answer is contained in the ideas that make up the question.

Before beginning, a note about the explanandum: When we ask 'why ought we ϕ?' there are two things we might be asking. We might want an explanation of why we ought, *all things considered*, to ϕ, or we might want an explanation of why we ought, *other things being equal*, to ϕ.[3] When it comes to why we ought to

[2] Clifford seems to be led to think that the constraint is a moral one because having beliefs not grounded in evidence can have morally grave effects, e.g. the destruction of a country. Of course, having beliefs not grounded in evidence can also have morally beneficial effects. A prisoner might, for example, adopt religious beliefs without evidence and do many more good things than he would have otherwise. So, it seems that if the promotion of moral action is the end for which we ought to believe, Clifford's conclusion is unwarranted.

[3] The same distinction can be put in terms of *pro tanto* and all things considered reasons. I avoid talk of reasons here because, in the doxastic realm, 'reasons' talk often gets conflated with 'evidence' talk in a way that will be unhelpful to this discussion.

believe the truth, some claim we have only an other-things-being-equal obligation to do so. I'll avoid that dispute here. In general, what we ought to do all things considered is fixed by what we ought to do other things being equal. Here I'll offer an explanation of the more foundational notion – why we ought, other things being equal, to believe the truth.[4] Proponents of an all-things-considered obligation to believe the truth can then build their explanation out of mine.

3.1 Beliefs Ought to be True

The account I'll give takes the concept BELIEF to be normative. The idea is often put like this: it's a conceptual truth that a belief that *P* is *correct* if and only if *P*.[5] It's important that the sense of 'correct' being used is normative, like in 'correct action.' This normative sense is distinguished from the merely descriptive deflationary sense of the term, where it means the same as 'true', as in 'correct supposition'. Following Ewing (1939), we can say that the *correct* belief, on the normative view, is the belief we ought to have, for some sense of 'ought.'

It is also important to distinguish the normativity claim from the ambiguous claim that belief *aims at the truth*. There are two senses in which belief might aim at the truth. According to the descriptive sense, belief is (generally) produced, revised, and destroyed by truth-conducive processes. The normative sense of 'aiming at the truth' is different: it says that beliefs *ought* to be true (independently of whether they are). Velleman (2000, p. 17) endorses both senses, but the account I give here will only require the normative version, which I'll defend as a conceptual truth:

NORMATIVE TRUTH REGULATION It's a conceptual truth about belief that if some attitude is a belief that *P*, one ought, other things being equal, have that

[4] As such, in the following explanation, I'll use 'ought' without restriction to mean the same as 'ought, others things being equal.'

[5] For example, see Shah (2003), Wedgwood (2002), and Shah and Velleman (2005). In the first paper, Shah only advocates for the 'only if' direction of this claim. In the later paper, the biconditional is advanced.

Along with each of these authors, I treat the concept BELIEF as normative because it generates a normative conceptual truth. Some authors, including the literature inspired by Quine (1951), Williamson (2003), and Williamson (2007, ch. 4), suspect that there are no conceptual truths. It's possible to endorse a normative account of BELIEF without thinking this requires it being a conceptual truth, but that story will not fit with the explanation I give below. In order to explain the phenomenon I discuss below, that belief is correct when true must be accessible to all knowledge attributers. I flesh out this accessibility in terms of it being a conceptual truth about belief. One option here is to treat conceptual truths as *rationally* accessible to agents who posses the concepts. As Williamson (2003, p. 291) notes, this kind of conception of conceptual truth avoids his criticisms there, and it also appears to avoid his attacks in Williamson (2007, ch. 4), as noted by Wedgwood (2007b, p. 18-21), which responds to earlier versions of those arguments in Williamson (2006).

attitude if and only if P.

There is a lot of disagreement in the literature about what is required for something to be a conceptual truth (see note 5). Here I will assume that if it is a conceptual truth about something x that everything that is an x has some feature, then nothing counts as an x unless it has that feature. So, in this case, NORMATIVE TRUTH REGULATION requires that if an attitude counts as a belief that P, then it is the sort of thing one ought to have if and only if P. Below, I will paraphrase this requirement with the phrase 'belief ought to be true'. Further, I will assume that conceptual truths are accessible, in some sense, to those who have the concept, so that concept holders only properly apply the concept to things that have the relevant feature.[6]

Notice that NORMATIVE TRUTH REGULATION is distinct from claims about norms of *justified* or *rational* belief. If a reliabilist view of justification is right, then in order to have *justified* beliefs, one's beliefs must be formed by a reliable process. Nonetheless, a belief may satisfy the proposed norm of mere belief (by being true) without being reliably formed. NORMATIVE TRUTH REGULATION is about a norm that governs belief itself, not justified or rational belief.[7]

Shah (2003) and Shah and Velleman (2005) defend a version of NORMATIVE TRUTH REGULATION. They argue that the concept BELIEF being normative in this way best explains the phenomenon of *transparency*, that "when asking oneself *whether to believe that p*, [one must] immediately recognize that this question is settled by, and only by, answering the question *whether p is true*" (2003, p. 447). Steglich-Petersen (2006) argues that a more sophisticated non-normative account of BELIEF can explain transparency equally well.[8] So, while I accept the argument from Shah and Velleman, I'll provide some new support for NORMATIVE TRUTH REGULATION.

3.1.1 New Support for the Normativity of Belief

John Greco (2003) argues that knowledge attributions function to give credit to the subject for their true belief. A lucky gambler may claim that she *knew* that her horse would win while her friend denies it. When they're fighting about the knowledge

[6] As I mention in note 5, in light of the arguments from Williamson regarding epistemic conceptions of conceptual truth, it might make sense to flesh out this kind of accessibility as *rational* accessibility, which is compatible with the conditions given here.

[7] Notice that this leaves open the possibility that the norms of belief spring from a deeper norm that all belief ought to be justified. My claim is merely that they are conceptually distinct.

[8] Shah and Velleman's argument also requires that transparency necessarily obtains. Its necessity isn't entirely obvious though, as Chappell (2005) points out.

attribution, what they seem to be disagreeing about is whether the gambler is *due credit* for her true belief that the horse would win. We also give credit to students (sometimes in two senses) for their true beliefs when they know the right answer to a question. If we discover that a student is not due credit for his true belief, perhaps because he cheated, we'd say he didn't know the answer in the first place. What seems to be going on here, as Greco claims, is that our knowledge attributions convey that the subject is due credit for his or her true belief. I'll call this 'the credit-giving phenomenon.' Greco argues that it can explain both the lottery and Gettier problems. In Gettier cases, subjects have true belief only by accident, so they are not due the credit they would receive from a knowledge attribution, Greco claims. The lottery problem is treated similarly.

Importantly, the the credit-giving phenomenon is distinct from the more controversial claims of virtue epistemologists, who claim that knowledge is, at least in part, a kind of credit-worthy state, like Zagzebski (1996), Sosa (2007), and Greco (2009). One may accept that this phenomenon obtains while still offering a traditional justified, true belief account of knowledge, for example, by merely claiming that credit-worthiness is entailed by having knowledge.

Suppose for a moment that Greco is right and that we do use knowledge ascriptions to give credit for true belief, at least in some important cases. Notice that giving credit for something requires having a positive evaluation of it. If I give you credit for being well-read in metaethics, I must think that there is something good about your being so well-read. On a first pass, this is the difference between giving you *credit* for it and *blaming* you for it: were I to think that your being well-read was negative, I would blame you for it instead.[9] Admittedly, some cases are tough to capture in this framework: We can credit Hitler for being the most ruthless dictator in history. Below, I argue that these outlier cases can be made sense of neatly, so for now, let's suppose that we give credit for something only if we view it in a positive light.[10]

[9] There is another sense of 'credit' that we use to attribute casual responsibility to people that doesn't require the positive spin. For example, I can credit an error in baseball without taking an error to be positive. The kind of credit in play in the phenomenon to be explained isn't this kind though, since of course, the gambler's true belief was causally his own.

[10] I'm being purposefully vague about what exactly this positive normative perspective is. One might flesh it out as viewing the thing as good, or at least having a complicated good-based attitude toward the thing, as Moore (1903, §1-2) would be inclined to do. Others might take the attitude to be one of viewing it as correct (right, positive), without reducing correctness (rightness, positiveness) to some kind of goodness. All I require here is that it's a positive normative assessment. Greco (2003, p. 121) takes being deserving of credit to require that the object have some kind of value, and he attributes this account to a natural widening of a view he attributes to Joel Feinberg (see Greco's fn

So, given our supposition that we use knowledge ascriptions to give credit for true belief, we must also positively assess true belief when we say that someone knows. This is the weaker phenomenon of POSITIVE EVALUATION OF TRUE BELIEF:

POSITIVE EVALUATION OF TRUE BELIEF When we attribute knowledge, we positively evaluate the subject's true belief.[11]

POSITIVE EVALUATION OF TRUE BELIEF is a weaker phenomenon than Greco's credit-giving phenomenon in that one can accept that POSITIVE EVALUATION OF TRUE BELIEF occurs without thinking that we give *credit* to subjects for their true belief when we say that they know. In fact, advocates of many accounts of knowledge very different from Greco's will also expect POSITIVE EVALUATION OF TRUE BELIEF.

Consider an account of knowledge that takes knowledge to be safely true belief, inspired by the kinds of views discussed by Williamson (2000) and Pritchard (2007).[12] The key idea of this kind of account is that knowledge is non-lucky true belief. So, they will hold that someone knows that *P* when their belief is true and free from error in all nearby possible scenarios.[13]

Intuitively, it seems like thinking of a mental state as knowledge requires taking it to deserve a kind of positive evaluation not due to mere true opinion. How might an advocate of the safe-true-belief account of knowledge explain this? A natural move here would be to appeal to the goodness of safety. The safety of the mental state (partially) adds the positive evaluation due to knowledge states that is not due to mere true opinion, the advocate would say.

Roughly, this is right, I think. Having a safely true belief is better than having a mere true belief, in the same kind of way that having knowledge is better than having mere true opinion. The betterness added by safety can't be due to the safety itself though, since safety by itself seems to be evaluatively neutral: Tripping a

14). My partial account of these positive judgments will be compatible with many ways of telling the whole story.

[11] There is a *de dicto* and a *de re* reading of this statement of the claim, but I intend to use only the *de dicto* reading, that we give credit for the true belief as true belief. This reading is supported by the credit-giving phenomenon, since credit-giving is intensional.

[12] Williamson (2000, ch. 4-7) spends a lot of a time discussing a (non-reductive) understanding of knowledge as requiring safely true belief. Williamson doesn't claim that this account is complete, so there may be other conditions necessary for knowledge. Williamson also doesn't offer any contenders here though, so the view is at least worth considering as complete. See Goldman (2010) for more discussion of this view. Pritchard (2007) also does not take safe true belief to be sufficient for knowledge, but he does take safety to be the core of a proper account of knowledge.

[13] This is an intentional simplification of the views actually on offer here. It will be easy to see why an advocate of this simplified view would expect POSITIVE EVALUATION OF TRUE BELIEF, and the main thread will carry over to the more sophisticated views.

friend while running a race is bad, for example, but *safely*, i.e. non-accidentally, tripping a friend is worse. Safety, then, isn't always a good thing. It seems to act like an evaluative magnifier: safely doing well is better than merely doing it, but safely doing badly is worse.

Once we see that safety acts like an evaluative magnifier, the advocate of the safe-true-belief account of knowledge needs a new story about why thinking of a mental state as knowledge is to think it deserves a positive evaluation not due to mere true belief. The addition of safety to true belief magnifies the positive evaluation due to true belief, so while true belief is due some positive evaluation, knowledge is due strictly more.

Notice though that in giving this account, the advocate of this account must admit POSITIVE EVALUATION OF TRUE BELIEF. In order for safety to magnify the positive evaluation due to true belief in the way required to make sense of our special evaluation of knowledge, the true belief must be taken to be positively evaluable in the first place.

One might worry that the story I've given here about the safe-true-belief account of knowledge being committed to POSITIVE EVALUATION OF TRUE BELIEF is peculiar because it relies on there not being some fourth aspect of knowledge that could be responsible for the distinct positive evaluation due to knowledge that's not due to mere true opinion. So, for example, an account of knowledge that takes it to be justified, safely true belief may appear not to be committed to the phenomenon occurring, since it could appeal to a distinctive value of justification.

Strictly speaking, this worry is correct; if the positive evaluation that's associated with knowledge is *totally* due to our positive evaluation of justification, in the example account, then its advocate need not be committed to POSITIVE EVALUATION OF TRUE BELIEF. She could just think that we can explain our positive evaluation of knowledge in terms of our evaluation of justification. This seems like an implausible claim though, and our exploration of the safe-true-belief account brings this out: all that's required to be committed to POSITIVE EVALUATION OF TRUE BELIEF is that we think that our positive evaluation of knowledge is *in part* derivative on our taking true belief to be positive. But of course knowledge attributors think that the value they ascribe to the knower is *partially* due to their true belief being valuable. So, thinking that there's a fourth aspect of knowledge *could* allow an account to escape thinking that POSITIVE EVALUATION OF TRUE BELIEF obtains but only at the cost of losing the ability to explain our positive evaluation of knowledge in terms of a positive evaluation of true belief at all. So any plausible story about knowledge attributions seems like it should be committed to POSITIVE EVALUATION OF TRUE BELIEF:

POSITIVE EVALUATION OF TRUE BELIEF When we attribute knowledge, we positively evaluate the subject's true belief.

Here, I'll assume POSITIVE EVALUATION OF TRUE BELIEF and show that NORMATIVE TRUTH REGULATION, the claim that BELIEF is a normative concept, best explains this phenomenon.

3.1.1.1 Explaining the Positive Evaluation of True Belief

Knowledge attributors positively evaluate the subject's true belief when they attribute knowledge. Any good account of this phenomenon must explain two aspects of it: it must explain in virtue of what knowledge attributors take true belief to be positive. Is it true belief's guidance value? Or, maybe, is true belief intrinsically positive? I'll call this the "metaphysical explanandum." Any good account of the phenomenon must also explain how knowledge attributors are in a position to be aware of true belief being positive. Even if true belief is valuable for it's ability to guide us, for example, how do knowledge attributors know that, especially in cases of abstract or esoteric true belief? This is the "epistemological explanandum." I'll argue here that taking BELIEF to be a normative concept best explains both of these explananda.

To see this, start by dividing the possible explanations of the phenomenon by how they respond to the metaphysical explanandum. If true belief deserves a positive evaluation, that positive evaluation must be due to something either intrinsic or extrinsic to true belief. I'll claim that just taking true belief to be extrinsically positive doesn't give an adequate explanation of the epistemological explanandum. Taking true belief to be intrinsically positive but not as a conceptual matter is subject to the same worry. That it's a conceptual truth that other things being equal, beliefs ought to be true offers the most plausible explanation of the phenomenon.

Many possible accounts of POSITIVE EVALUATION OF TRUE BELIEF take true belief to be positive merely in virtue of its relationship to something else. True belief might be good, for example, insomuch as it has *guidance value*, the ability to help us complete our projects. Certainly, for example, true beliefs about the location of the cookie jar will help the cookie monster in his quest.

But, POSITIVE EVALUATION OF TRUE BELIEF can't be explained solely in terms of true belief having guidance value. If it were, we'd expect knowledge attributors to take true belief to have guidance value at least most of the time. It's hard to think we usually do this. As Goldman (1999, p. 3) points out, "The dinosaur extinction fascinates us, although knowing its cause would have no material impact on our lives." Similarly, it's hard to see how a set theorist's knowledge about transfinite ordinals or the librarian's knowledge of the first word of the 171st page of *Ulysses* is guiding. Of course, one might be able to contrive strange situations in which these esoteric bits of knowledge might guide us, but generally, we don't think of these knowledge contents as having guidance value.[14] Knowledge attributors simply don't need to take a subject's true belief to be guiding to ascribe knowledge. So, that knowledge attributors positively assess true belief because we take it to have guidance value seems implausible.

In trying to locate the source of epistemic normativity in our desires, Hilary Kornblith (1993) gives a sophisticated argument to the conclusion that true belief is universally instrumentally valuable for the achievement of goals. The value of true belief, on this account, is derived from the value we place in satisfying our goals.[15] Following this line, we might expect that we could explain the knowledge attributor's positive assessment of true belief in terms of its universal instrumental value.

Kornblith's account, even if it is right, doesn't fully explain how knowledge attributors would positively assess true belief though. As I mentioned above, in order to explain POSITIVE EVALUATION OF TRUE BELIEF, we need more than just an account of what true belief is taken to be positive for. We also need an explanation of how knowledge attributors might be in a position to assess true belief as positive in the way described. This is the demand of the epistemological explanandum. If knowledge attributors don't know about true belief being universally instrumentally valuable as Kornblith claims, even if it is valuable as such, knowledge attributors wouldn't view it that way. Kornblith's account doesn't naturally lend itself to an account of this, the epistemological explanandum.

One might think that Kornblith's account could be extended to answer the epistemological explanandum; it would say that knowledge attributors view true beliefs as universally instrumentally valuable. But, this seems implausible. Consider again the knowledge attributions about transfinite cardinals. It's hard to imagine

[14] For example, knowledge about transfinite cardinals might be guiding in some very odd betting situations. We don't typically attribute knowledge of advanced mathematics with this kind of guidance value in mind though.

[15] Kornblith takes one of the neat features of his argument to be that true belief turns out to be instrumentally valuable for the achievement of our goals, regardless of what those goals are.

that we give the set theorist credit for his true belief because we take it to be pragmatically valuable (even if it is valuable in that way).[16] So, even if Kornblith's account is right about the value of true belief, the account doesn't adequately explain POSITIVE EVALUATION OF TRUE BELIEF.

We can generalize the worry above to most accounts of POSITIVE EVALUATION OF TRUE BELIEF that take the positive assessment of true belief to be about true belief's relationship to something else: Let *e* an extrinsic end, like guidance value, for which true belief is purportedly positively assessed by knowledge attributors. In general, if it's possible for there to be someone who lacks a concept of *e* and still attributes knowledge, that knowledge attributor can't positively assess true belief because of its relation to *e* when she attributes knowledge, since she can't even think thoughts of *e*. For many things *e* extrinsic to true belief, it seems possible for there to be such a knowledge attributor. For example, Albert is a possible brain in a vat who can only think thoughts about set theory and knowledge. Albert, it seems, should be able to attribute knowledge of a mathematical theorem to himself even if he lacks concepts required to think about guidance value. So, true belief's relationship to guidance value can't be for what knowledge attributors positively assess true belief. This model is easily generalizable into an argument schema that can be used to show against most accounts of POSITIVE EVALUATION OF TRUE BELIEF that attempt to explain the metaphysical explanandum in terms of something extrinsic to true belief.

Of course, there are some possible accounts of the phenomenon in terms of something extrinsic to true belief that aren't subject to the kind of argument given. If an account of the phenomenon can explain the positive evaluation in terms of true belief being viewed as positive in virtue of its relationship to something extrinsic, the end has to be something that knowledge attributors must have a concept of. Such accounts include those that take justification or knowledge to be in virtue of what true belief is positively evaluated. I'll argue later that these accounts are also lacking. Before that though, let's consider accounts that try to explain POSITIVE EVALUATION OF TRUE BELIEF in terms of true belief being intrinsically positive.

Goldman, among many others, takes it that "true belief is the ultimate value in the epistemic sphere" (2002, p. 53). This view, *veritism*, takes true belief to be valuable for its own sake.[17] It might seem well-suited to explain how knowledge

[16] Again here we can strengthen the case by considering a knower whose intentions are wholly evil, like the terrorist case above. There, a knowledge attributor wouldn't positively assess true belief for being instrumentally valuable to that agent.

[17] See Grimm (2009) for a discussion of veritism.

attributors positively assess true belief, since it locates the source of the value of true belief in something tangible to any knowledge attributor, namely the true belief itself.

Mere veritism doesn't tell the whole story though; even if true belief is intrinsically valuable, an agent cannot give credit for it unless he also familiar with that value. To explain the phenomenon, we need an explanation of how knowledge attributors are acquainted with the intrinsic value of true belief – that is, veritism must also answer the epistemological explanandum. Veritism tells us that true belief *is* valuable, but it doesn't explain how knowledge attributors know about that value. It must be supplemented with an epistemology of the intrinsic value – one that makes the value of true belief accessible to knowledge attributors.

NORMATIVE TRUTH REGULATION can offer both parts of the explanation. There's a simple story about the positive evaluation of true belief: According to NORMATIVE TRUTH REGULATION, it's a conceptual truth about belief that belief ought to be true. So, true belief is belief the way it ought to be.[18] Unlike veritism, NORMATIVE TRUTH REGULATION can also account for the epistemological explanandum. Since it is conceptual truth about belief that beliefs ought to be true, we have a simple story about how knowledge attributors are aware of that: having the concept BELIEF is required for thinking KNOWLEDGE thoughts in the first place, and having the concept BELIEF is sufficient for having access to the requirement that belief ought to be true.

So, NORMATIVE TRUTH REGULATION offers a simple explanation of both the metaphysical and epistemological explananda. As such, the proposal offers a better explanation of the credit-giving phenomenon than any of the other kinds of accounts considered so far.

One might think that if we took normative epistemology to be *a priori* (but not conceptual), like Gibbard (1990, 2003), Audi (2005), Huemer (2005), and Wedgwood (2007a, p. 245), we could explain the phenomenon equally well. Its being *a priori* that beliefs ought to be true, according to these accounts, explains how it is accessible to all knowledge attributors.

The *a prioristic* accounts aren't so well-off, though. If these accounts also allow for genuine normative disagreement, they cannot explain the credit-giving phenomenon as well as NORMATIVE TRUTH REGULATION: Tooley, for example, thinks that abortion is permissible and Marquis disagrees. Their disagreement is not about

[18] In this way, NORMATIVE TRUTH REGULATION agrees with veritism. NORMATIVE TRUTH REGULATION says strictly more than veritism by making it a conceptual truth about belief.

any non-normative matter of fact. It's about the distinctly normative issue of the permissibility of abortion. *A prioristic* accounts like Gibbard's (2003, esp. ch. 7 and ch. 14) go through great pains to explain how this kind of genuine disagreement is possible. But if disagreement about the distinctly normative is possible on these views, then it should be possible for a knowledge attributor to mistakenly not assess true belief positively, if we take it to be merely *a priori* that beliefs ought to be true. The accounts don't explain how, these knowledge attributors, the ones for whom it is *a priori* that belief ought to be true but who fail to realize it, positively evaluate true belief.[19]

To finish canvassing the possible explanations of POSITIVE EVALUATION OF TRUE BELIEF, let's return to the kinds of accounts I deferred above that take the positive evaluation of true belief to be due to its relation to justification or knowledge. Like my account, these possible accounts can also locate the source of the proposed goodness of true belief in things accessible to knowledge attributors, namely that the true belief is justified (assuming that knowledge requires justification) or that the true belief is knowledge. As such, they are not subject to the general argument schema from above.[20] But, as explanatory theses, these proposals are less satisfying than NORMATIVE TRUTH REGULATION. Here's why:

First let's consider accounts of POSITIVE EVALUATION OF TRUE BELIEF that attempt to explain the positive assessment of true belief in terms of its relationship to justification. If justification is not required for knowledge, like in the account of knowledge offered by Lewis (1996) and the safe-true-belief kinds of accounts discussed above, then the account will be subject to the general argument schema given above. There will be a possible knowledge attributor who positively assess the subject's true belief when she attributes knowledge, but she couldn't be doing that in virtue of true belief's relationship to justification because she lacks the relevant concept. So, let's consider how the justification story fares when justification is taken to be required for knowledge.[21]

[19] This problem can be taken as a dilemma: either the accounts claim that these agents must see true belief as good (in the relevant sense) or they allow that they may not. If agents must always view true belief as good, then they must explain why this normative judgment is unlike others, where we can disagree. If they allow that agents may not, they are still saddled with the project of explaining how these agents can make knowledge attributions (or hold, counterintuitively, that they cannot). By treating the concept BELIEF as containing the normative standard, we can explain more generally how knowledge attributors are in a position to give credit for true belief.

[20] I'll grant that knowledge attributors have access to the proposed value or goodness of knowledge or being justified.

[21] At first glance, taking our positive evaluation of true belief to be due to its connection to justification may appear to be too strong. It seems to get the Gettier and lottery problem cases

One benefit of thinking that the true belief of knowers receives a positive evaluation by the attributor is that it opens the door to an easy account of the Gettier and lottery cases. This is a generalization of Greco's claim that we can solve these problems by thinking that knowledge claims give credit to the subject for their true belief. On Greco's account, the positive evaluation is a particular type – the kind associated with giving credit. But a story similar to Greco's will go through about the Gettier and lottery cases as long as we take the positive evaluation of true belief to be a kind of positive evaluation that wouldn't be due to a Gettiered or lotteried subject.

Accounts that appeal to true belief's connection to justification to explain POSITIVE EVALUATION OF TRUE BELIEF lose the ability to tell this kind of story about Gettier and lottery cases. This is because if the positive evaluation of true belief that knowledge attributors have is due to true belief's connection to justification, we should expect that if the subject is sufficiently secure in her justification and the subject's belief bears the right connection to that justification, the subject should be due that positive evaluation. In the Gettier and lottery cases though, there are no special restrictions on the connection between the subject's belief and her justification, and subjects can bear as strong a justificatory relationship to the believed true proposition as needed to suggest that the positive evaluation is due. So the justification-based accounts of POSITIVE EVALUATION OF TRUE BELIEF ought to predict that in some Gettier and lottery cases we do give the positive evaluation of the true belief. But then, those accounts lose the ability to explain why we withhold knowledge attributions in those cases by appeal to the lack of the positive evaluation of the subject's true belief.

For example, consider a justification-based account that takes the type of positive evaluation knowledge attributors give to true belief to be the credit-worthiness kind, like Greco's. Roughly, according to such a view, justification is the source of the

wrong. In those cases, the agents *do* have justified (true) belief, but we don't give give the positive evaluation. So if we explain the positive evaluation of true belief in terms of justification, the account seems to predict that we should give the thumbs up to those cases.

That argument against the justification view is too fast though. If it were correct, it would show against my view as well, since Gettier and lottery subjects also have true beliefs. It's important to notice that, according to the justification view, it is a necessary, but not sufficient, condition for receiving the positive evaluation that one have justification. There could be other conditions that are also required for the evaluation. For example, it might be that being due the positive evaluation requires that the justified true belief also spring from the subject's own abilities, i.e. the subject is responsible for it. In the Gettier and lottery cases, agents have the justified true belief, but it fails to spring from their own abilities. So, they are not due the positive evaluation, according to both my view and the proposed justification view.

credit-worthiness judgment, and true belief is valuable as a means to justification. So the account ought to predict that the agent is due the positive evaluation for her true belief when she is due credit for her *justification* and her true belief bears the right relation to that justification. In the Gettier cases though, the subject goes through a competent deduction to arrive at her justified true belief. So, *justification* does spring from the agent's abilities, and she *is* due credit for her justification. So, this view would predict that we do give the positive evaluation in the Gettier cases, and therefore, it wouldn't be able to explain the Gettier cases as Greco aims to. A generalized version of this reasoning shows that other justification-based accounts of POSITIVE EVALUATION OF TRUE BELIEF also cannot employ the simple story about the Gettier and lottery cases. NORMATIVE TRUTH REGULATION avoids this reasoning by locating the source of the positive evaluation of true belief in the true belief itself, not in its relation to anything else.[22]

Another type of account that avoids the general argument from above is one that explains POSITIVE EVALUATION OF TRUE BELIEF in terms of true belief's relationship with knowledge, which they take to be the only relevant thing of value. One might think that Williamson's (2000) primacy of knowledge argument naturally motivates this kind of view. The idea being that since the concept KNOWLEDGE doesn't breakdown cleanly into any other concepts, when we positively assess some mental state as knowledge, we don't take the true belief to be a valuable proper part of knowledge. Rather, we're positively evaluating knowledge, and the true belief that constitutes that knowledge comes along for the ride.[23]

This kind of knowledge-based account would most naturally reject the phenomenon being explained here. Were an advocate of such an account to accept the phenomenon, she would be claiming that though knowledge attributors see

[22] One might worry that this means that NORMATIVE TRUTH REGULATION predicts that the positive evaluation (that is supposed to explain the Gettier and lottery cases) must be present anytime true belief is. This doesn't follow. NORMATIVE TRUTH REGULATION only makes true belief a necessary condition for the positive evaluation. Other natural conditions will also be required for the positive evaluation, like perhaps that the agent is due credit for that true belief, as Greco requires.

[23] Though this kind of reasoning may motivate some readers, I don't think it ought to. The primacy argument for knowledge that Williamson (2000) gives are arguments for the primacy of analysis, not primacy of value. The primacy of analysis of KNOWLEDGE gives us no reason to think that what we take to be positive about knowledge is irreducible to what we take to be positive about some of the parts. Suppose, for example, that GENOCIDE is analytically prime – that is, that genocide cannot be understood solely in other terms (as say, many killings combined with some political motivation). Even if that's so, we may still think that the badness of genocide is partially due to the badness of the killings. Similarly for knowledge: even if KNOWLEDGE is unanalyzable, that doesn't give us reason to think that the positive evaluation associated with it doesn't ultimately come from the positive evaluation we give to true belief.

knowledge as the source of the positive evaluation, we none-the-less positively evaluate true belief when we say that someone knows. But, if we think that the knowledge itself is the source of the positive evaluation, why would we think that we're positively evaluating the true belief (rather than the knowledge) with knowledge attributions? This is an unnatural position to hold. Since my claim here is that NORMATIVE TRUTH REGULATION offers the most plausible explanation of POSITIVE EVALUATION OF TRUE BELIEF and these knowledge-based accounts ought to just reject the phenomenon rather than explain it, I will ignore the knowledge-based accounts here.[24]

According to NORMATIVE TRUTH REGULATION, it's part of the idea of belief that true belief is belief the way it ought to be. I claim that this best explains why when we attribute knowledge, we positively evaluate the subject's true belief. An alternative proposal might claim that the phenomenon is best explained by a symmetrical fact about TRUTH – that the concept TRUTH is normative. Such a proposal seems to be able to employ all of the same moves I do in defense of my claim about BELIEF, due to truth's similar positioning with belief in the phenomenon. Price (1998) attributes to Wright (1994), for example, the claim that "Any reason to believe that p is true is a reason to believe (and hence allow the assertion) that p" is a norm of truth. Many others have made similar claims about norms of truth that relate belief and assertion to truth.[25]

I agree with many of the proposed norms and in particular the claim that a proposition being true makes it the case that a belief in that proposition is correct. That said, these proposed "norms of truth" seem to be better understood as norms of belief or assertion than as norms generated by an inherent normativity of the concept TRUTH. If a concept is normative, we should expect that in general when the concept applies, some 'ought' claim or prescription follows. If something is *wrong*, for example, it ought not be done, or if something is *disgusting*, one ought to avoid

[24] Notice that there is worry for the knowledge-based accounts though when they are combined with Greco's claims about the nature of the positive evaluation as credit-giving: Recall Greco's claim that the credit-giving phenomenon can explain the Gettier and lottery problem cases. There, our unwillingness to give credit is supposed to explain our unwillingness to attribute knowledge. If that's right, then explanations of the credit-giving phenomenon that take true belief to be good in service of *knowledge* leave us with an unsatisfyingly small explanatory circle: the Gettiered subject is not due credit for his true belief because his belief does not constitute knowledge, and his belief does not constitute knowledge because he is not due credit for it. While there is no inconsistency in this position, it's not very enlightening either.

[25] Price (1998, p. 248) claims, for example, that it's a norm of truth that "One is incorrect to assert that p if, in fact, it is not the case that p." Horwich (2006, p. 347) also describes a similar norm of truth: "It is desirable to believe what is true and only what is true." For more examples, see Engel (2002).

it.[26] TRUTH doesn't seem to be like that. Take, for example, a paradigmatically non-normative claim, like "snow is white." Since "snow is white" is paradigmatically non-normative, I take it that it doesn't entail any (non-trivial) 'ought' claims or prescriptions. If TRUTH is a normative concept, we should expect that "snow is white is true" entails some 'ought' claims or prescriptions. But "snow is white" and "snow is white is true" seem to entail exactly the same claims. So, that's one way in which TRUTH doesn't appear to work like other normative concepts.[27] Taking BELIEF to be a normative concept is a more plausible explanation of NORMATIVE TRUTH REGULATION than taking TRUTH to be normative.

In the preceding discussion, I canvassed many of the possible explanations of POSITIVE EVALUATION OF TRUE BELIEF. The considerations there show that NORMATIVE TRUTH REGULATION makes sense of the phenomenon in ways that the other views cannot. The views that attempt to explain the metaphysical explanandum in terms of something extrinsic to true belief generally fail to explain the epistemological explanandum, and the same worry applies to the view that attempts the explanation in terms of the intrinsic goodness of true belief. Those extrinsic accounts that may be able to explain the epistemological explanandum, the views given in terms of true belief's connection to justification and knowledge, don't explain the phenomenon very well either. The justification-based views lose a natural story about the role of the positive evaluation in explaining Gettier and lottery cases, and the most natural versions of the knowledge-based accounts reject the phenomenon to be explained. Finally, I argued that a view that takes TRUTH to be a normative concept (rather than BELIEF) is less plausibly correct than NORMATIVE TRUTH REGULATION, since TRUTH doesn't seem to exhibit the symptoms of normativity. While I haven't surveyed *every* possible explanation POSITIVE EVALUATION OF TRUE BELIEF, the above considerations are sufficient to provide a new reason to think that NORMATIVE TRUTH REGULATION is true. Taking the concept BELIEF to be normative gives us a simple story about both for what and how knowledge

[26] We might think there are trivially normative concepts, i.e. concepts that are understood in normative terms but need not be. For example, we could generate a concept NORMRED that applies to all things that are red and such that if murder is wrong, we ought not murder. Such a concept is normative, in some sense, because to capture the meaning of the term, one must use normative language. But this new concept is not normative in a standard sense where normative concepts non-trivially guide. Surely, the advocates of the claim that TRUTH is normative, insomuch as it's an interesting claim, must take it that TRUTH is normative in a non-trivial way.

[27] One might also think that because 'true' factors in true norms, this is enough to show that it's a normative notion. I don't think this is right. For example, it's a norm of chess that one ought to checkmate the opponent. But what it is for the opponent to be checkmated is just for one of a number of possible board positions to occur, which can be understood completely non-normatively.

attributors positively evaluate true belief. It's a conceptual truth that true belief is belief the way it ought to be, and because this is part of the very idea of belief, that fact is accessible to knowledge attributors.

Above, when I motivated thinking that POSITIVE EVALUATION OF TRUE BELIEF occurs in the context of Greco's conception of knowledge ascriptions, I claimed that when we give credit for something, we must take that thing to be, in a loose sense, positive. But there are some hard cases for this claim: We can credit Hitler for being the most ruthless dictator in history, and being a ruthless dictator isn't positive. Before concluding this section, I'll show that these hard cases are resolved once we clarify the kinds of things we give credit for.

The explanation of POSITIVE EVALUATION OF TRUE BELIEF I endorse says that we positively evaluate true belief because it's a conceptual truth about belief that belief ought to be true. This is an instance of a broader pattern of positively evaluating satisfying conceptually-given standards, I claim. For example, it's a conceptual truth about playing chess that players ought to follow the rules, and we give a player credit for playing chess correctly when he does. Mark Schroeder (2010) makes a similar point with this example: We need not approve at all of being an assassin, but nonetheless, we can make sense of what it is to be good *by the standards of being an assassin*. It's being an effective killer. That's a conceptual truth about being an assassin. We can give credit to the effective killer for being good *as an assassin*. When we credit Hitler, we're applauding him for satisfying the conceptually-given standards *of the activity in which he was engaged,* being a ruthless dictator. This is how we give credit in the hard cases: we give credit for satisfying the standards of an activity. Similarly for belief: we positively evaluate true belief because true belief satisfies the standard that's part of the idea of belief, that of being true.

3.2 Agents as Believers

The truth norm is supposed to be categorical – we all ought to believe the truth, regardless of the particular projects or aims each of us has. If NORMATIVE TRUTH REGULATION is true, where does this leave us on the project of explaining why we ought to believe the truth? According to the proposal, it's part of the concept BELIEF that beliefs ought to be true. So, to think of an attitude as belief, rather than as supposition, desire, or some other attitude, in part requires thinking that it ought to be true.[28] But, just showing that a concept has a normative aspect doesn't show that we're subject to that constraint, just as showing that something is a requirement of etiquette doesn't show that we're subject to it either. So we need a further story about why that norm of belief applies to us.

What we do get from the claim that BELIEF is normative in the sense of NORMATIVE TRUTH REGULATION is that *if someone is a believer*, then he or she ought to believe the truth.[29] The charge then is to show that the antecedent is satisfied by every agent. That is, we must show that every agent is a believer. Then we'd have a complete explanation of the norm:

(1) NORMATIVE TRUTH REGULATION: It's a conceptual truth about belief that if some attitude is a belief that P, one ought, other things being equal, have that attitude if and only if P.
(2) So, if S is a believer, S ought to believe the truth.
(3) All agents are believers.
(4) So, agents ought to believe the truth.

In this section, I'll argue that claim (3) is true: agents are believers, in that they have beliefs in the normative sense defended above.

It is a presupposition of action theory that an *agent* is an entity that performs actions. What then is required to perform an action? On the belief/desire model, an agent's actions can be explained in terms of her beliefs and desires. It is Jon's desire for ice cream, for example, that when combined with his belief about the location of the ice cream, explains why he went to the kitchen. Donald Davidson's (1963) highly influential account offers roughly this picture. There, one's actions are explained in terms of, among other things, a *primary reason*, which is the pairing of a desire and an instrumental belief of the agent. On this picture, agents must be believers. Many other philosophers follow Davidson in requiring agents to have similar kinds of beliefs.

The belief-desire-intention model of agency (Bratman 1987), which has become a standard model of agency in artificial intelligence research (Georgeff et al. 1999), also models agents in terms of their pro-attitudes and beliefs.

Modern decision theories and formal epistemologies, such as Bayesianism,

[28] Probably this isn't all that's required to think of an attitude as belief. For one, this doesn't separate belief from hypothesis, which is plausibly also normatively truth regulated. Belief, for example, also seems to be non-normatively regulated by truth, as explained by Shah and Velleman (2005).

[29] Again, here, as above, the sense of 'ought' is the restricted other-things-being-equal sense.

also require rational agents to have a kind of belief. In these theories, agents are represented as having partial, degreed beliefs. The degreed beliefs then factor essentially into agents' decisions and consequent actions.[30]

The common theme is that in order to be an agent, an entity that acts, one must also be a believer.[31] So, agents must be believers. When we combine this with the conceptual truth about belief, that if S is a believer, S ought to believe the truth, we get the desired categorical norm: All agents ought to believe the truth.

One might think that we can explain the actions of agents in a way that avoids the normative implications by appealing to other attitudes of the agent. But, the possibility of developing this kind of account of action is daunting, I'll claim. We'd expect there to be some attitude – call it 'schmelief' – that plays the kind of role belief was meant to play in standard explanations of action. When we try to characterize schmelief, we should restrict it to being an attitude that agents actually have, since we can't appeal to merely potential attitudes to explain actual actions. Further, we'd expect that if schmelief can play the role of belief in agents' reasons, it must be an attitude that represents its content as true. That is, we'd expect schmelief to be a cognitive, rather than conative or affective attitude. Examples of non-belief attitudes like this include supposing, imagining, assuming, hypothesizing, remembering, accepting, judging, and understanding.

Many of these options for schmelief require that agents believe in order to have them. Understanding, for example, seems to require believing: For José to understand that $2 + 2 = 4$, he must believe it as well. Remembering, accepting, and judging are like this, too. So, schmelief can't be one of these attitudes, if we're aiming to avoid the normative implications of NORMATIVE TRUTH REGULATION.[32]

Attitudes that are generally insensitive to evidence, like supposition, don't seem well-suited to be schmelief either. Say that Jon wants ice cream and merely supposes that there's ice cream in the kitchen. Jon then enters the kitchen. When we want to know what Jon's reason was for moving to the kitchen, we can't point to his supposition. That's partly because he would have had that purported reason even

[30] Treating belief as degreed or graded is not problematic for the view that belief ought to be true, it just requires us to reformulate what it means for those beliefs to be true. See Joyce (2009) for further discussion of this issue.

[31] Agents having beliefs might also be required to explain the identity of agents over time, in terms of their memories or expectations, Railton (1997) points out in discussing this kind of account.

[32] One might think that schmelief could be *acceptance*, like the notion used in the constructive empiricism of Van Fraassen (1980, p. 12). According to that view though, "acceptance of a theory involves as belief only that it is empirically adequate." Because accepting that P requires having a belief, accepting cannot be the non-normative notion we're looking for here.

if his evidence about the location of the ice cream were widely different. In general, to explain an action, we must show why the action would have made sense from the agent's perspective. But, Jon knows that his suppositions are insensitive to his evidence, so he wouldn't rely on that attitude to guide his action.[33] This kind of worry would apply to other proposed explanations of action in terms of generally evidence-insensitive attitudes, like imagining and assuming. So, schmelief must be an attitude that is generally tied to evidence in the right way. Otherwise schmelief can't explain how actions make sense to agents.

So we see that for schmelief to play the right role in explaining actions, it must look a lot like belief.[34] A last resort option here would be to claim that schmelief just is the same attitude (property) as belief, but the concept SCHMELIEF provides a non-normative way of describing that property.[35] Employing this move doesn't avoid my conclusion though: The hedonist can admit that the concept MAXIMIZING NET PLEASURE picks out the good in a non-normative way, without thereby denying that we ought to do good things. Analogously, if SCHMELIEF is a non-normative way of picking out belief, this doesn't show against the normativity of belief being able to explain the categoricity of the truth norm.[36]

Given that schmelief, insomuch as it can play the belief role in explaining action,

[33] Even if Jon's supposition is right and he knew it, the reason for Jon's movement couldn't be his *supposition* (along with his desire). That's because Jon's supposition is still insensitive to the evidence. In other nearby possibilities where Jon doesn't believe that the ice cream is in the kitchen, the supposition remains.

[34] My opponent might try to appeal to a disjunctive attitude to avoid my claim here. For example, if we let schmelief be an attitude that ought to be true in regard to propositions that might guide action, such as ones about locations of cookie jars, but the same attitude doesn't put any normative constraints on propositions that cannot guide us, perhaps such as ones about large cardinals. There are three problems with this kind of move: First, we have no reason to think agents have such an attitude, so positing it seems to be more costly overall than accepting my conclusion. Second, phenomenologically, it doesn't seem that how we regard propositions differently based on their potential guidance to us, especially since we often lack insight into whether a proposition will potentially guide us. Thirdly, such a story would still accept my explanation of why we ought to believe the truth about a wide variety of propositions, namely the potentially guiding ones. So, such an account doesn't really avoid my conclusion in an interesting way.

[35] Gibbard (1992) argues that, in general, we cannot decompose thick normative concepts (like what I am taking BELIEF to be) into their normative and non-normative components, as this move would require. McDowell (1981, p. 144) offers a similar argument. For the sake of strengthening my opponent's position, I will ignore those arguments here.

[36] More explicitly, the hedonist will accept the following argument: (1) Goodness ought to be promoted for its own sake. (Conceptual Truth); (2) Goodness is pleasure. (Claim of the hedonist); So, (3) pleasure ought to be promoted for its own sake. Analogously, the advocate of explaining action in terms of schmelief (thought of as the same property as belief) ought to accept the following argument: (1) Belief ought to be true. (Conceptual Truth); (2) Belief is schmelief.; So (3) Schmelief ought to be true. The conclusion of this argument won't be a conceptual truth, but it can still explain why the truth norm applies categorically.

must be an evidence-sensitive cognitive attitude that isn't the same property as belief, it's difficult to see exactly what attitude it would be. It's much more plausible that when agents act, it's their *beliefs* explain their actions. But to think of that attitude as a *belief* is to think that it's an attitude that ought to be true.

At the beginning of the discussion, I set out to discuss the source of some epistemic norms. Clifford points at a particular norm about evidence and claims that the norm applies to us in the same way that moral norms apply to us. My goal here was to provide an alternative account, not of Clifford's norm, but of the truth norm, a norm that seems to be more fundamental than Clifford's norm. According to the truth norm, we must believe the truth. So, to provide an explanation of that norm, I must show how it could arise and in virtue of what it applies. I claim to have done that. The explanation I provide is this:

(1) NORMATIVE TRUTH REGULATION: It's a conceptual truth about belief that if some attitude is a belief that P, one ought, other things being equal, have that attitude if and only if P.
(2) So, if S is a believer, S ought to believe the truth.
(3) All agents are believers.
(4) So, agents ought to believe the truth.

This give us an alternative to Clifford's understanding of epistemic normativity: we can understand (at least some) epistemic norms as having a non-moral, distinctly doxastic, source. It's part of the very idea of belief that beliefs ought to be true, and because agents must all believe, that norm applies to all agents.

3.3 Some Objections and Replies

Now that I've set out a distinctly doxastic way of understanding the truth norm, I'll consider some objections to it as an explanation of that norm.

Objection: More and More Beliefs?

One might think that the account I offered above, if right, would require agents to constantly be searching for new truths to believe. This would require agents to be constantly counting blades of grass and memorizing phone books. And even if the account does not require that they must do these things, since the normative requirement on agents is only an other-things-being-equal constraint, the account seems to suggest that agents always have *some* reason to do these things, the objector claims.

While some authors do think that we always have some reason to adopt new true beliefs, perhaps because having a true belief is intrinsically valuable, the account I offered is neutral on this debate. The account I offer here only requires agents to have some beliefs, namely beliefs about those things that might factor into the explanations of the agent's actions. The normative claim generated only puts a constraint on those beliefs that agents have by requiring agents to have that belief if and only if the content is true. When agents do not have a belief regarding *P*, my account is silent. It's open then to proponents of views that require us to gather more and more beliefs to explain why we must adopt beliefs in those cases.

One might worry then that the explanation I give is too weak, in that it doesn't require agents to have enough true beliefs. Can't the norm generated by my account just be satisfied by not believing or reducing the number of beliefs we have to a minimum?

As I argued above, agents cannot just stop believing *in toto*. It's a requirement of being an agent that one have beliefs, at least enough to explain one's actions. Further, a descriptive feature of belief is that it is often difficult to fail to have a belief about many topics. If, for example, you believe the snow is white, it's typically not open to you to stop having a doxastic attitude about that topic. Even when you do change your attitude by suspending belief or coming to disbelieve the thing we previously believed, like when you're presented with new evidence about an issue, you're still subject to the norm. To disbelieve *P* is just to have a belief that not-*P*, so this new belief is subject to the norm. Further, we treat an agent's suspension of belief as a kind of belief in explaining action, like when a driver slows down at an unmarked intersection when she's unsure if someone is coming. So, suspension of belief, like disbelief, counts as a kind of belief in the explanation I offer. So given that the agent must adopt some doxastic attitude, the norm generated would say that the agent ought, in the relevant sense, to stop suspending belief and believe that *P* iff *P*.

Objection: What kind of 'ought'?

According to the account that I offer, we must believe the truth because there's an 'ought' that's part of the concept of belief. We might wonder why that ought applies to a believer just in virtue of being a believer. To see the objection, consider an analogy to chess: it's plausibly part of the idea of chess that when playing chess, one ought move the bishop only diagonally. But, you might think that there could be reason for someone not to move the bishop only diagonally when playing chess, such as if she would suffer a large financial loss by doing so. So, in the case of chess, it seems that it might not be that we ought to move the bishop only diagonally even if that norm is part of the very idea of chess. The same worry is supposed to apply to the account I offer. The objector rejects the second claim in my explanation: it's not the case that if someone is a believer, he or she ought to believe the truth.

There are two important aspects of the account I'm advancing that collectively reply to this objection. Firstly, belief is disanalogous to chess in that it we aren't essentially chess players. What I claim to have shown in the previous section is that it's essential to being an agent that we believe. So while it could be the case that we have all-things-considered reason to stop playing chess (such as if we would suffer a large financial loss by continuing to play), as agents, we can't opt out of being believers. The natural picture here then is that we're subject to the norms of an activity when we're participating in that activity.[37] When we would suffer a large financial loss by moving the bishop only diagonally, we have good reason to stop participating in playing chess (though we may still have reason to *appear* to be playing chess). In the case of belief, insomuch as we ought to do anything, we're agents, so we're stuck being believers. If we're believers, the norms of belief apply to us.

Secondly, the above considerations may make it sound like I'm committed to thinking that we can never have all-things-considered reason to believe something false. This result doesn't follow from my account. All that follows from my account is that *other things being equal* we ought to believe the truth. So, believers are always subject to this other-things-being-equal 'ought'. Unlike in chess, where it is never permissible to move the bishop except diagonally, the demand on agent's beliefs according to my account is only that agents have some reason to believe the truth. This is enough to explain the categorical nature of the norm but not so strong as to anticipate that agents may only believe the truth.

Together, these two considerations show that the objection by analogy to chess is not apt. We're subject to the norms of the activities in which we're engaged. Unlike chess, we cannot opt out of being believers. Also unlike some of the norms of chess, the truth norm that's constitutive of belief doesn't require that agents have

[37] Participating in an activity need not be active or intentional. Consider, for example, those unwitting German citizens who participated, though intentionally, in the atrocities of the Second World War. The notion I'm using here is similar to, but strictly more general than, Rawls's (1955) notion of a practice.

an overriding obligation to believe the truth. This flexibility explains how agents can engage in activities that put contravening demands on the agent's attitudes.

Objection: But Why Should I *Believe the Truth?*

The account of the truth norm that I offer is a kind of constitutivist account, like those of Korsgaard (2009) and Velleman (2000). The account is constitutivist because it locates the source of the norm for belief in what is constitutive of belief. Constitutivist accounts of norms share the goal of giving a non-hypothetical account of why we must conform to (epistemic, in my case) norms in terms of something constitutive of agency or action. There's a popular kind of objection to these kinds of accounts. Enoch (2006, p. 169), quoting David Lewis, puts the objection like this:

> Why care about objective value or ethical reality? The sanction is that if you do not, your inner states will fail to deserve folk theoretical names. Not a threat that will strike terror into the hearts of the wicked! But whoever thought that philosophy could replace the hangman?

The worry here is straightforward: Constitutive accounts of norms claim that norms are grounded in our agency, but a skeptic may just reply that she's not interested in agency (or whatever other notion the constitutivist appeals to). 'What reason to I have to be an agent, rather than a schmagent?', the skeptic will ask.

This worry is quite compelling against constitutivist accounts that aim to respond to skeptic of this kind. But, responding to this kind of skeptic is beyond what is required to give an account of epistemic norms. Consider an analogy to accounts of medical disorders: what we expect of a good account of a medical disorder is that it tell us the conditions under which and in virtue of what one has the disorder. Good accounts of medical disorder need not make it undeniable (or even accessible) to the patient that he or she has the disorder. In this sense, theories of medical disorders need not respond to the skeptic. We should expect the same of accounts of norms. To give an account of a norm, it is sufficient to give the conditions under which and in virtue of what the norm applies. Theories of norms, like theories in the sciences, need not make the application of the properties they discuss accessible to the things to which they apply.

My goal in this chapter was to offer an account of why we ought to believe the truth that doesn't treat it as a moral or pragmatic issue. To do this, I needed to give the conditions under which and in virtue of what the norm applies, and I've done that. It's part of the very idea of belief that beliefs ought to be true. Being an agent requires being a believer. So, agents ought to believe the truth; the norm applies to all agents in virtue of their agency.

CHAPTER IV

Epistemic Rationality and the Aim of Belief

ABSTRACT

In "Rational Credence and the Value of Truth," Allan Gibbard (2008c) makes two key claims about the nature of epistemic rationality: He claims (1) that epistemic rationality cannot be explained just by the aim of belief, and (2) that epistemic rationality is more plausibly explained in terms of a practical notion, that of maximizing prospective guidance value. In this chapter, I reconstruct and respond to Gibbard's arguments for both of these claims.

In order to understand Gibbard's second claim, we must clarify the role of ideal theory and idealizations in the explanation Gibbard proposes. When we do, we see that Gibbard is left appealing to merely hypothetical bets to explain epistemic rationality even in ideal cases. These merely hypothetical bets don't seem to be able to play the role he needs them to though. I show that Gibbard's story is also unique in this respect; he can't appeal to the kinds of explanations given to save Dutch Book arguments. So Gibbard must explain how these explanations work.

I then undermine Gibbard's motivation for pursuing such an account in the first place, by showing that the argument for the first claim is unsound. Gibbard isn't using the best conception of the aim of belief. I survey five senses in which something or someone can be said to 'aim.' According to the normative sense in which 'belief aims,' belief aims at the truth because having a true content is the standard of good or correct belief. By taking belief to have an aim in this way, I sketch how we can make sense of epistemic rationality in terms of that aim.

When we evaluate other people, we distinguish between someone being epistemically rational and someone being practically rational.[1] Broadly speaking, an agent exhibits epistemic rationality when she properly regulates her beliefs in light of the information available to her, and she exhibits practical rationality when she pursues her goals the best she can by her own lights.

A natural way to distinguish epistemic and pragmatic rationality is in terms of the aims of the corresponding projects. Whereas practical projects aim at the fulfilment of one's goals, epistemic projects aim at the production of true belief and the elimination of false belief. In "Rational Credence and the Value of Truth," Allan Gibbard (2008c) poses two worries for this way of distinguishing these types of rationality. The first worry is that epistemic rationality cannot be fully understood just in terms of the pursuit of truth. This is because exhibiting epistemic rationality requires a kind of immodesty that is not guaranteed just by valuing true belief. The second worry is that the aspect of epistemic rationality that cannot be explained by the pursuit of truth can be explained in terms of a more practical aim. Being epistemically rational, Gibbard then concludes, should be understood not in terms of the value of true belief but instead in terms of potential guidance value.

Gibbard makes many idealizing assumptions about the agents under consideration, what kinds of situations they face, and what epistemic rationality requires of them. Gibbard appeals to a formal result to show that for Bayesain agents in a specific betting scenario, epistemic rationality is best understood in terms of the agent aiming at maximizing the guidance value of her beliefs. In this chapter, I'll respond to Gibbard's argument. In the first part of the response, I'll argue that Gibbard's use of idealization is a peculiar in ways that leave his explanation of epistemic rationality in need of further exposition. Gibbard's account is best conceived of as employing hypothetical collections of bets in explanations, I show. I then suggest that the practical requirements generated by these hypothetical bets don't seem to be able to underpin the epistemic explanada in the way that Gibbard requires.

Gibbard was motivated to pursue such a view in the first place because, as he argues in the first part of his article, he believes that epistemic rationality cannot be understood in terms of the aim of belief. In the second part of my response, I undermine Gibbard's motivation for pursuing his practical account of epistemic rationality in the first place. I show that Gibbard takes too narrow a view on what

[1] Other authors use the terms "theoretical rationality" and "instrumental rationality" to talk about what I'm calling "epistemic rationality" and "practical rationality" respectively. I don't want to take a stand here on where those notion are the exact same kind of rationality as the ones I talk about here, but they certainly are very similar.

the aim of belief could be. By conceiving of the aim of belief in a popular way that is different from Gibbard's conception, I'll sketch how we can recover a picture of epistemic rationality grounded in that aim.

4.1 Gibbard's Argument

Gibbard's argument is quite dense and integrated, though I think it is best understood in two parts, one part arguing for each of the two worries I ascribed to Gibbard above. The first part is meant to show epistemic rationality cannot be fully understood in terms of agents valuing the truth. In the second part, Gibbard claims that epistemic rationality is more plausibly understandable in terms of belief aiming at maximizing guidance value.

Following Gibbard, throughout the chapter, I'll think of the beliefs of agents as representable by a probability function, which represents their degrees of credence in the propositions they have (partial) belief about.[2] What an agent desires and values will be represented with a utility function on possible states of the world to how much the agent desires or values that state. Then, following Ramsey (1926), I'll assume (again with Gibbard) that what agents ought to decide to do is what maximizes their expected utility relative to their credences and utility assignments.

4.1.1 Gibbard's Argument Part I: Epistemic Rationality and the Value of Truth

"Belief aims at the truth" is often given as a gloss of the normative or functional connection between belief and truth. This aim of belief is thought by some to be able to explain why we ought to regulate our beliefs for truth, rather than for pleasure or something else we value. If that's right, it seems plausible that other norms that apply to belief, such as norms of justified, rational, or warranted belief (or even knowledge) could be made sense of in terms of this deeper connection between belief and truth.[3]

Gibbard starts with this intuition and tries to unpack the idea that belief aims at truth, which he takes to be mostly metaphorical: "Belief . . . can't aim literally; it's we who aim" Gibbard (2008c, p. 143) tells us. So, ideal believers must be representable

[2] Thinking of the doxastic states of agents as representable by a probability function grants to agents a lot of what others have claimed is required by epistemic rationality. Gibbard (2008a) acknowledges this point but intends to restrict his result to only these idealized agents. My arguments do not question this idealizing assumption.

[3] Along with Wedgwood (2002), I think something like this picture is right. Unfortunately, fully defending this conception of epistemic norms is beyond the scope of what is possible here.

as aiming for truth in their beliefs, Gibbard concludes. Gibbard takes this idea to underpin a minimal requirement for epistemic rationality. For an agent to be epistemically rational, she must aim at the truth in her credences. More precisely, if an agent is epistemically rational, she must satisfy this minimal test for epistemic rationality:

MINIMAL TEST FOR EPISTEMIC RATIONALITY "When a person forms her credences with epistemic rationality, is it as if she were choosing her credences voluntarily, rationally aiming, in light of her credences, at truth in those very credences" (2008c, p. 147).

Intuitively, to think to a person as epistemically rational, Gibbard is suggesting, we must think of that person as by their own lights pursuing having accurate credences.

The first of Gibbard's two claims is that epistemic rationality cannot be understood solely as resulting from an agent purely *valuing* the truth of their beliefs.[4] This is because an agent purely valuing the truth in her beliefs doesn't guarantee that she'll pass the MINIMAL TEST FOR EPISTEMIC RATIONALITY. Here's why:

An agent purely values the truth in her beliefs when the agent prefers her beliefs to be as accurate as possible – that is, when the agent's utility function orders possible situations where she has credences in terms of how accurate they are in that possibility.[5] Gibbard formalizes this idea for a simple toy agent who only has beliefs about one proposition, that modern Europeans descend at least in part from the Neanderthals, S: Let $g_1(x)$ be how much utility the agent assigns to having credence x in S if S is true and $g_0(x)$ be how much utility the agent assigns to having credence x in S if S is false. Then, according to Gibbard, pursuing the truth, or purely valuing the truth in one's beliefs, consists in satisfying CONDITION T:

CONDITION T Function $g_1(x)$ increases strictly monotonically with x, and function $g_0(x)$ decreases strictly monotonically with x.

In words, the agent purely values truth when she prefers having higher credences in S if S is true, and she prefers having lower credences in S if S is false (Gibbard

[4] It's important to note that the claim here is that epistemic rationality cannot be made sense of in terms of *valuing* the truth. The claim, here at least, is not that it can't be made sense of in terms of *aiming* at the truth, since (what Gibbard takes to be) the metaphor of aiming is underwriting what Gibbard takes to be the minimal standard of epistemic rationality in the first place. Below, I'll have more to say about Gibbard's argument and *aiming* at the truth.

[5] Strictly speaking, Gibbard puts this in terms of the agent being representable as such, not that the agent must really have such a utility function. For ease of presentation, I'll say the agent 'has' such a utility function when the agent is representable as such.

2008c, p. 148-9). More generally, the intuition here is that when an agent values some goal, she takes the states of the world in which the goal is achieved to be preferable to those where it is not achieved.

Can valuing the truth, in this sense, be all there is to being epistemically rational? Gibbard doesn't think so. Being epistemically rational requires more, Gibbard claims. In order to pass the MINIMAL TEST FOR EPISTEMIC RATIONALITY, epistemically rational agents must also be representable as having utility functions that are credence-eliciting:

CREDENCE-ELICITING A utility function is *credence-eliciting* when, in light of some credences Cr, the expected utility of adopting Cr is higher than the expected utility of adopting any other credence function.[6]

That is, in light of their evidence, epistemically rational agents, prefer to have the credences that they do have in light of that evidence.

For sufficiently ideal agents, a requirement to be as if one has a credence-eliciting utility function is generated just by the MINIMAL TEST FOR EPISTEMIC RATIONALITY. Here's why: Suppose (for reductio) that there were an epistemically rational ideal agent who, in light of her evidence, has a credence of .5 in S, and suppose that, in light of her evidence, the agent prefers, as part of her concern for truth, to have a credence of .7 in S. Then the agent can rationally advance her quest for the truth by changing her .5 credence to .7. So, the agent does not view her credences as optimal in the pursuit of her goal of having accurate credences. According to the MINIMAL TEST FOR EPISTEMIC RATIONALITY though, for the epistemically rational agent, it is as if she rationally chooses her credences with the goal of maximizing the accuracy of those credences, but our agent is not like that. So, epistemically rational agents must have credence-eliciting utility functions.

Are there ways then to purely value the truth then without satisfying the MINIMAL TEST FOR EPISTEMIC RATIONALITY? Yes. This is because there are utility functions that satisfy CONDITION T but are not credence-eliciting. The linear score is an example: $g_1(x) = x$ and $g_0(x) = 1 - x$. If an agent adopts this linear score as her utility function, then she satisfies CONDITION T because $g_1(x)$ increases strictly monotonically with x, and function $g_0(x)$ decreases strictly monotonically with x. So, we can think of this agent as purely valuing the truth in her beliefs. But unless her credence is 0, .5, or 1, she prefers to have a credence that she doesn't have. If

[6] When the utility function views the credences as *at least as good* as any other, then the utility function is *proper*. If a utility function has views the credences as strictly the best, then it is strictly proper. By "credence-eliciting," Gibbard means strictly proper. See his footnote 11 on p. 152.

her credence is $> .5$ then she views having credence 1 as maximizing her payoff in terms of believing the truth, and if her credence is $< .5$, she views credence 0 as maximizing that payoff. If the agent is purely concerned with the truth then, by her own lights, she can *rationally* advance that concern my changing her non-extreme to an extreme one. But this kind of move would be "epistemically rash," so merely valuing the truth cannot be all there is to being epistemically rational, Gibbard (2008c, p. 152) concludes.

Gibbard's argument has a number of moving parts, but we can understand it by analogy to a much more tangible example. Suppose that I want to make some guacamole to accompany dinner tonight. There are a number of things that are required of me if I am to rationally pursue that goal. The question that is analogous to the one that Gibbard pursue is whether we can make sense of what is rationally required for me to do in pursuit of guacamole just in terms of my goal of having guacamole.

So suppose I am purely concerned with making guacamole for dinner. An agent is purely concerned with making guacamole when the agent's desires are representable by a certain class of utility functions, those that prefer states of the world in which one makes guacamole. So, to *purely value* making guacamole is to satisfy CONDITION G:

CONDITION G The utility function of the agent orders states of the world in which the agent makes guacamole before those in which she doesn't.

Further, in order for an agent to be rational (practically or instrumentally, in this case) with respect to the concern for making guacamole, the agent must obtain avocados, since they're required to make guacamole. More generally, practical rationality requires that agents *pursues necessary means*:

PURSUING NECESSARY MEANS An agent *pursues necessary means* with respect to a goal g when, if ϕ-ing is a necessary means for attaining that goal, the agent tries to ϕ.

Clearly, being practically rational requires that agents satisfy PURSUING NECESSARY MEANS, but the requirements of PURSUING NECESSARY MEANS with respect to making guacamole (i.e. getting avocados) are not guaranteed by merely by having a pure concern for making guacamole. So, in this practical case, being practically rational with respect to the guacamole goal cannot be reduced to valuing that goal.

The epistemic case differs from this simple practical case in that the requirement of practical rationality is a practical requirement, rather than an epistemic one.[7] Epistemic rationality, per Gibbard, requires credence-eliciting utility functions in the same way that practical rationality requires pursuing necessary means. Neither of these requirements of rationality is guaranteed by an agent adopting the corresponding values. So, Gibbard concludes, a "pure concern with truth for its own sake [cannot] explain epistemic rationality" (2008c, p. 159).

4.1.2 *Gibbard's Argument Part II: Epistemic Rationality as Maximizing Guidance Value*

In the first part of the argument, Gibbard argues that epistemic rationality cannot be made sense of in terms of a pure concern for truth. The reason why is that an agent can prefer accuracy in her credences without without thereby satisfying all of the requirements of epistemic rationality – in particular, without having a credence-eliciting utility function. In the second part of the argument (the way I'm reformulating it), Gibbard argues that epistemic rationality is more plausibly reducible to aiming at maximizing expected guidance value. Requiring that agents maximize the expected payoffs of their beliefs, unlike requiring that they value the truth in their beliefs, does require the agent's utility function to be credence-eliciting.

To see this, notice that an agent's utility function being credence-eliciting is a property of the functions $g_1(x)$, how much utility the agent assigns to having credence x in S when S is true, and $g_0(x)$, how much utility the agent assigns to having credence x in S when S is false. What makes a pair of functions credence-eliciting is that the expected utility, when calculated from the perspective of the agent, of having the credence that the agent actually has is higher than the expected utility of having any other credence. Formally, a pair of functions $g_1(x)$ and $g_0(x)$ is credence-eliciting just in case they stand in a very particular relationship to each other. Gibbard (2008c, p. 152-3) discusses the particular way this works, but the formal aspects of that won't concern us here. For our purposes, we can consider a particular pair of functions as an example: If we let $g_1(x) = -(1-x)^2$ and $g_0(x) = -x^2$, we get the Brier score, a well-studied credence-eliciting utility function. Using this pair of utility functions, we can calculate the utility of having some credence c in S is $-(1-x)^2 c + -x^2(1-c)$. This quantity is maximized

[7] There is another important difference the reader should notice: In the case Gibbard actually considers, he is granting to the opponent that the agent in question maximizes her expected utility. He then goes on to argue that *even if the agent is rational in these other ways*, that doesn't guarantee his epistemic rationality. In the practical case, we should make that same concession since much more plausibly, maximizing expected utility *is* all that's required for practical rationality.

when $x = c$, so if the agent's credence were c and she uses this pair of utility functions to calculate the expected accuracy of some credence x, she'd view c, the credence she actually has, as maximizing that expected accuracy. Of course, there are many (uncountably-many) more pairs of credence-eliciting pairs of utility functions. Following Gibbard, I'll call such a pair of functions *SAM-qualifying*.

What Gibbard points out is that the collection of SAM-qualifying pairs of utility functions is exactly the same as the collection of pairs of utility functions such that if an agent has those utility functions, she views her own credences as maximizing expected guidance value. From this, Gibbard concludes that since epistemic rationality requires agents to be as if they have credence-eliciting utility functions, and this is exactly what is guaranteed by the agent trying to maximize the expected *guidance* value of their credences, it is this maximizing of expected guidance value that best explains epistemic rationality.

To get at these ideas, Gibbard offers an example of a person trying to avoid a tiger that is behind one of two doors, either the left or the right, one of which he must open. If the tiger is in fact behind the right door, then choosing the right door to open has a utility of -100, let's say, and similarly for the left door. We'll say that the *guidance value* of the policy to open the right door is -100. Gibbard does not give a precise characterization of what constitutes guidance value, but he does give some prototypes: In discussing the possible valuable aspects of beliefs, he says, "Beliefs can be comforting. They can be empowering. They can link one to others in a fellowship of conviction. I'll label all the kinds of value that credences can have apart from their guidance value as *side value*" (2008c, p. 155). Guidance value, we might try saying, is the value a belief or policy has in helping us promote our goals, separate from any goals we might have that involve that belief itself.

Suppose that the tiger is in fact behind the right door and that the agent has a credence of .4 that the tiger is behind the right door. Then the *actual* guidance value of his credence is -100. That is, if the agent follows through on the actions that maximize his expected utility, then he'll get -100 utility. The agent cannot tell that this is the actual guidance value of his credence though, since he doesn't know the location of the tiger. In general, the actual guidance value of an array of credences will depend on factors not accessible to the agent. We can ask what the agent can *expect* the guidance value of his credences to be though. In this case, from the agent's perspective, it appears that the value of opening the right door is -40, because he has a .6 credence that the -100 value possibility is on the left side. This is what Gibbard calls the "prospective guidance value" of the credence (2008c, p. 156).

When it comes to prospective guidance value, practical rationality seems to require that we adopt those credences that have the highest prospective guidance value by our own lights. When we adopt such a policy, we're being "guidance-immodest." A way of choosing credences with respect to a utility function is *guidance-immodest* when "if one forms one's credences that way, one will attribute to credences formed in that very way, then, a maximal expectation of guidance value" (2008c, p. 156). What Gibbard shows, using a formal result of Schervish (1989), is that we maximize the prospective guidance value of our credences on certain a series of bets iff our utility functions are credence-eliciting. It is this practical, guidance-seeking aim that seems to underpin our seemingly epistemic obligation to take ourselves to be maximizing accuracy in our credences then, Gibbard concludes.

To see how this works, recall our agent who is trying to pick the most accurate credence she can for S. Gibbard proposes that we can offer her a series of bets $\mathfrak{G}_\beta$ for $0 < \beta < 1$ on whether S such that the payoff of bet G_β is $\overline{\beta} : \beta$. If an agent has a credence x in S, then she'll view all and only the bets G_β such that $\beta > x$ as better than fair (and she'll accept them using standard decision theory). What credence would an agent wish herself to have then if she aims to maximize her expect payoff for this series of bets? She expects she can do no better than by acting on the credence she actually has, since having any other credence would have her accept some bets she now takes to be unfair or reject some she now takes to be better than fair. So, in this case, the agent is guidance-immodest.

Gibbard then appeals to the formal result from Schervish (1989). Here is Gibbard's restatement of the result:

> Smooth functions g_1 and g_0 are credence-eliciting if and only if for some possible continuum of bet offers and a policy of accepting any bet offer G_γ exactly when $\gamma < x$, $g_1(x)$ gives the expected payoff of the policy given S, and $g_0(x)$ gives the expected payoff of the policy given $\overline{S}$.[8]

The series of bets $\mathfrak{G}_\beta$ that Gibbard gives is a series that satisfies the right-hand side of the biconditional, so we can conclude that the utility functions g_1 and g_0 of the agent are credence-eliciting.

So, by constructing the series of bets on S, Gibbard shows that the agent having a credence-eliciting utility function, though it is not demanded by having a pure concern for truth, is required if the agent's credences maximize their expected

[8] This is Gibbard's (2008c, p. 157) restatement of a formal result gestured at by Schervish (1989, p. 1869). It ignores some minor formal issues that will be irrelevant to the discussion here.

guidance value for those bets. Since epistemic rationality requires being as if one has credence-eliciting utility functions, Gibbard claims, a pure concern for truth doesn't seem well suited to explain epistemic rationality (that's part one of the argument), but a concern for maximizing the expected guidance value of one's credences can explain it (that's part two). As Gibbard goes on mention, since the value of true belief is only a side value, it cannot factor into the guidance value, the prospective maximization of which seems to explain the requirement of epistemic rationality that our concern for truth be SAM-qualifying. Gibbard puts the point like this:

> Simply wanting truth or accuracy for its own sake does not explain [why our concern for truth must be SAM-qualifying]. Wanting truth entirely for the sake of guidance would explain it – and this is the only explanation we have found. (2008c, p. 160)

Putting it all together then, if we aim at the truth by valuing accuracy in our credences, the aim of truth can't explain epistemic rationality, Gibbard argues. Epistemic rationality requires that we aim at the truth in a special way – a way that lines up with aiming to maximize the prospective guidance value of our credences. It is this second aim then that best explains epistemic rationality, Gibbard suggests.

4.2 Saving the Epistemic from the Pragmatic

As I presented it above, Gibbard offers a two-part argument in which he tries to show that epistemic rationality is best understood in practical terms. Gibbard's proposed practical understanding of epistemic rationality relies heavily on the use of idealized agents and situations: Gibbard argues that for a perfect Bayesian agent that can "[perform], at will, whatever acts have highest expected utility as reckoned using [her] credences" (2008c, p. 156), that agent will be epistemically rational when she's offered a certain infinite collection of bets. In this section, I'll claim that Gibbard's use of idealization is peculiar and that Gibbard must tell a further explanatory story. There are some possible avenues of response open to Gibbard here and canvassing all of them is not my main focus here. Instead, the worry I pose about Gibbard's idealization should give us reason to consider whether there are other options open for explaining epistemic rationality. In the second part of this chapter, I provide another such option. I show that Gibbard's motivation for rejecting the most natural type of explanation of epistemic rationality, in terms of the aim of belief, is unsound. He employs a bad understanding of the aim of belief. I'll survey various different kinds of aims there are and point out arguments in favor of interpreting the aim of belief radically differently from how Gibbard does. I then claim that we can understand epistemic rationality in terms of that a different understanding of the aim of belief and sketch the beginnings of such an account.

4.2.1 On Reducing Epistemic Rationality to Practical Rationality

In the second part of Gibbard's argument, he proposes that we explain the demands of epistemic rationality in terms of the demands of practical rationality. On this picture, our epistemic obligations derive from our pursuing what is best for us in terms of prospective guidance value.

It's important to notice that Gibbard's proposed explanation, in order to be plausible, must appeal to very idealized practical decision scenarios. His proposal can't be that the demands of epistemic rationality for *actual* agents are derivative upon a practical requirement to maximize the *actual* prospective guidance value of their beliefs. Here is a case that shows that:

NERVOUS BETTER Richard, a psychotherapist, is being shown the outcomes of flips of a weighted coin that is 55% likely to come up heads on any given flip. He starts out assuming that the coin is fair but then correctly comes to have a .55 credence that the coin will come up heads on the next flip after a very large number of trials. Richard also knows that he's about to be offered a series of bets $\mathfrak{H}_\beta$ for $0 < \beta < 1$ such that the payoff of bet H_β is $\overline{\beta} : \beta$ on whether the coin will come up heads on the next flip. Since Richard is a psychotherapist, he is very cognizant of how he acts under pressure and knows that when he's faced with bets, he tends to doubt his own mental arithmetic. Richard anticipates that when he's offered the bets, he'll probably doubt his views, assume the coin is closer to fair than he now takes it to be (since most coins are), and thereby not maximize his prospective returns. Richard knows that he can prevent this by making his credence more extreme, since he will then moderate his actions the appropriate amount when he's confronted with the bets. So, Richard actively tries to inculcate in himself a .6 credence that the coin will come up heads by seeking out evidence for that, imagining explanations of why the coin would be so biased, and asking friends to reinforce that idea.

In cases like this, maximizing the actual prospective guidance value of our beliefs in exactly the betting situation Gibbard describes seems to require us to be epistemically irrational. As a practical requirement, the real Richard must not accord his beliefs to his evidence if he is to maximize his prospective returns.

One might say in reply to this case that Gibbard is offering an ideal, rather than non-ideal, theory of epistemic rationality. The idea would be that Gibbard's account tells us about the 'epistemically rationally Utopian' state, where agents always act with full epistemic rationality and are not limited by the kinds of problems that face Richard above.[9] A full account of epistemic rationality would then supplement Gibbard's story with a more realistic story about limited agents.

This kind of response to the case is right but not quite complete. Gibbard *is* working in ideal theory, but we shouldn't think of Gibbard as offering a description of the ideal in presenting the betting scenarios.[10] When we do ideal theory for some realm, such as ideal political theory, we give a description of how things would ideally be in that regard, such as satisfying the difference principle, and then explain why they are that way, such as being the result of a system constructed by a fair procedure. When Gibbard offers his explanation of epistemic rationality in terms of the practical, he is already taking for granted the relevant features of the epistemically rationally ideal state, namely that the agents are Bayesian agents who can be represented as having a credence-eliciting utility function. Gibbard's aim is then to explain in virtue of what the ideal state is that way. The betting scenarios are supposed to explain the features of the epistemically rationally ideal world: in the ideal state, we're epistemically rational because we're practically rational with respect to this series of bets, Gibbard claims.

Idealization plays two distinct roles in this methodology: First, there is the difference between limited, everyday agents, such as you and me, and ideally epistemically rational agents, who are perfectly Bayesian, have credence-eliciting utility functions, and can "[perform], at will, whatever acts have highest expected utility as reckoned using [their] credences" (2008c, p. 156), among other perfections. This first kind of idealization is present in the methodology because we're doing ideal theory. After Gibbard restricts his attention to agents who are ideal in this first way, there is another kind of idealization – one that treats the ideal agents as only facing an infinite series of bets, having credences about only one proposition,

[9] The idea here is analogous to the ideal/non-ideal theory distinction typically made in political philosophy. See Valentini (2012) for an overview of the ways the distinction functions there.

[10] Suppose Gibbard were offering a description of the epistemically rationally ideal state in giving the betting scenarios. Then we would have to take Gibbard to be saying that in the epistemically rationally ideal world, agents are faced with infinite series of bets about each proposition they have credences about. Epistemic rationality certainly doesn't require that.

caring only about money, etc.[11] To keep the terminology clear, let's call the first kind of idealization, moving from limited to ideal agents (that factors in switching from non-ideal to ideal theory), *idealling*. Objects that are the result of idealling are *ideal*. I'll use *idealization* and *idealized* only for the second kind of idealization, which Gibbard uses in moving from talk about ideal agents in general to talk about ideal agents in ideal betting cases.

In these terms, we ought to view Gibbard as working entirely within ideal theory, so he is only talking about ideal agents and ideal scenarios. Gibbard then takes up the task of explaining the constraints of epistemic rationality on these ideal agents. He does this by ideal*izing* the ideal agent: he treats the ideal agent as being in an ideal betting scenario. For this idealized ideal agent, their being practically rational can explain their being epistemically rational, Gibbard claims. I will grant Gibbard's claim about the idealized ideal cases, but I'll ask what the import of these idealized cases is for the demands of epistemic rationality on non-idealized ideal agents, such as ideal versions of ourselves.[12] I suggest we look towards an understanding of the role of idealizations in explanations to find out. When we do, we'll see that it's unclear what Gibbard's idealizations have to say about the non-idealized ideal cases.

4.2.1.1 Concerning Idealization in Gibbard's Explanation

I take it that, generally speaking, we use idealized agents and situations in epistemology because doing so allows us to focus on the important explanatory features of the phenomena that we're interested in. Consider an analogy to science: we might try to understand the motion of a spring by treating it as an idealized simple harmonic oscillator, thereby restricting our focus to the major factors in the spring's motion. In creating an account of epistemic rationality, by intentionally idealizing away potentially distracting aspects of real cases, we can hopefully give a tractable account of epistemic rationality in terms of its core features. The goal then would be to apply the understanding gained from the idealized cases to unidealized cases.

Michael Weisberg (2007) catalogues three kinds of idealization: *Galilean idealization*, where distortions are intentionally introduced to make phenomena more easily

[11] Although supposing that the agents have a "sufficiently rich prospect for what 'bets' one will face in life" (2008c, p. 159) is the main idealization of the second type, the second type also includes many other idealizing assumptions, including those that I list here and those that Gibbard lists both in the concluding section of his article and in Gibbard (2008a).

[12] As such, I'll restrict the discussion below to discussion of ideal, though not necessarily idealized, agents.

tractable, *minimalist idealization*, where the phenomenon is idealized to include only its core explanatory features, and *multiple-models idealization*, where different incompatible models of the phenomenon are used to understand it.[13] The kind of idealization present in Gibbard's cases doesn't look like multiple-models idealization, since Gibbard is not offering different incompatible stories that explain various aspects of the epistemic rationality in the ideal agents. So, if the idealization in Gibbard's cases is one of these kinds, we should expect it to be either Galilean or minimalist idealization.

The question of whether the kind of idealization employed by Gibbard (and philosophers more generally in these kinds of argument) is Galilean or minimalist is not one I intend to answer here. Instead, it's enough for our purposes to notice that if the idealization is one of those two types, the main explanatory features of the idealized cases (or something quite similar to them) must be present in the rest of the ideal (though not ideal*ized*) cases to be explained. More precisely, both Galilean and minimalist idealizations satisfy SAVES MAJOR FACTORS:

SAVES MAJOR FACTORS If M* is an idealized version of M and $C_1, C_2, \ldots$ are the primary causal or explanatory elements in M*, then each of $C_1, C_2, \ldots$ or something that plays a similar role is present in M.

Intuitively, SAVES MAJOR FACTORS tells us that either the parts of the idealization that are important for explaining the idealized phenomenon or something that plays the same role in the explanation exists in the unidealized target phenomenon to be explained. When we offer explanations in terms of an idealized version of a target, we should want our idealizations to satisfy SAVES MAJOR FACTORS. If an idealization that we appeal to does not satisfy SAVES MAJOR FACTORS, explanations that may work in the idealized cases may not work for the unidealized case, since the explanatory parts of the idealization may be missing analogous parts in the unidealized case.

Minimalist idealizations satisfy SAVES MAJOR FACTORS straightforwardly: minimalist idealizations are idealizations which include only the primary explanatory features of the real phenomenon, so every primary explanatory feature must be part of the real phenomenon. Galilean idealizations are idealizations that leave out

[13] Weisberg's discussion takes place in the context of a discussion about idealization in scientific theories. I see no reason to think the lessons from that discussion don't apply to idealization in philosophy as well, with some natural generalization. For example, Weisberg characterizes minimalist idealizations as including only the core *causal* features; by generalizing that to the core *explanatory* features, we make the account more easily application to philosophical discussion where causation doesn't play the same central role as it does in the sciences.

or simply elements of the phenomenon to make it more computationally tractable. For example, we may idealize a spring by treating it as frictionless and a simple harmonic oscillator. Not every primary feature of a Galilean idealization is necessarily present in the real phenomenon, since some features may be computationally simpler approximations of the features of the real phenomenon, such as the simple harmonic oscillation in the spring example. In that case though, something similar to the feature of the idealized model will be present in the real phenomenon, namely the computationally hard real part of the phenomenon (dampened harmonic oscillation). So Galilean idealizations will also satisfy SAVES MAJOR FACTORS.

Gibbard's proposed explanation of epistemic rationality appeals to idealized versions of ideal agents. Gibbard's idealized cases do not seem to satisfy SAVES MAJOR FACTORS though. For example, consider ideal Sandra:

GRANDMA'S PHONE NUMBER Sandra's grandmother passed away about a month ago, and her family is working to settle her affairs. The family sold grandma's house and closed her account with the phone company months ago. Today, Sandra received a letter from the phone company confirming that the account had been closed and that the phone number would be reused no sooner than 100 years from now, so saving that number wouldn't be useful. Just below that, the letter lists the details of the closed account including the phone number. The last digit of the phone number is smudged though, and Sandra can't read it. Seeing that made Sandra wonder what that number was. Through the smudge, she can see that whatever number it was, it was curved on top. So, she lowers her credence that is was the number 4. Sandra then goes on dealing with her grandmother's estate. The question of her grandmother's phone number never comes up again and having had that number would not have helped Sandra in achieving any of her other goals.

In this unidealized case, epistemic rationality puts demands on Sandra that she satisfies: In lowering her credence that the last digit was 4, it seems that Sandra is acting epistemically rationally – coming to that conclusion is the proper response to her evidence. Because this is an instance of the phenomenon that Gibbard purports to explain, namely epistemic rationality in ideal agents, if Gibbard's idealization satisfies SAVES MAJOR FACTORS, we should expect that all of the primary explanatory features of Gibbard's idealized cases are present here. In the idealized cases, the infinite series of bets that plays the practical role is key to the practical understanding of epistemic rationality, but nothing seems to play a similar role in this unidealized case. There is no practical element to do the explanatory work GRANDMA'S PHONE NUMBER. So, it seems that Gibbard's idealizations do not satisfy SAVES MAJOR FACTORS.

Perhaps the kind of idealization that Gibbard uses is not one of the types of idealization given by Weisberg. Even if so, we'd still expect that kind of idealization to satisfy SAVES MAJOR FACTORS. That's because, as I mentioned above, explanations of a target phenomenon that appeal to elements of idealized models only seem to be able to explain the target insomuch as those elements exist in the unidealized cases. Explanatory features of the idealized case can't explain the unidealized cases, even by analogy, if those features don't exist in the explananda.

Gibbard might reply that the mere possibility of there being a practical element in the ideal cases is enough to explain the epistemic rationality. In GRANDMA'S PHONE NUMBER, for example, it's *possible* for Sandra to have been faced by the type of bets Gibbard's explanations require. This is where Gibbard's formal result pulls through: Using any credence-eliciting scoring rule, such as the Brier score, we can create a series of bets for which the agent would have the required prospective guidance value maximizing policy if the agent is epistemically rational. The proposal would then be that the series of bets that the agent faces is purely hypothetical. In all cases of epistemic rationality, there will always be this merely hypothetical series of bets that Gibbard could appeal to. If Gibbard were to then treat the bets as hypothetical also in the idealized cases, Gibbard's idealizations would satisfy SAVES MAJOR FACTORS.

I worry about a version of Gibbard's claim that treats the bets as merely hypothetical for both the idealized and unidealized ideal agent. Doing so removes the practical import of the bets from the agents. The practical import of the bets is crucial to the sort of explanation that Gibbard is offering of epistemic rationality in terms of the practical in the idealized case. To conceive of the sort of explanation that Gibbard is offering without the real gains and losses to the idealized agent is just to think of it as offering an account of epistemic rationality as *modelable* as a practical requirement on the agent. This wouldn't license Gibbard's conclusion that the explanation given is a practical one, rather than another type. For example, it's *possible* that the total utility in the world could have been directly tied to an ideal agent's doxastic life and maximized when the agent is epistemically rational. This mere possibility doesn't license the conclusion that the requirement that the agent be epistemically rational stems from moral obligation, even by the utilitarian's lights. He would have a moral obligation only if that possibility were actual. The analogous worry holds for the hypothetical version of Gibbard's claims: if the claim is just that there could have been a series of bets such that the idealized agent is practically rational with respect to them only if he is epistemically rational, this doesn't show that the agent being epistemically rational in that idealized case can be explained in practical terms.

So suppose then that Gibbard's explanation treats the bets as non-hypothetical for the idealized ideal agent. GRANDMA'S PHONE NUMBER shows us that, in order for Gibbard to explain epistemic rationality in all ideal cases, the explanation must still treat the bets as merely hypothetical for some unidealized ideal agents. The natural story then would be that hypothetical bets play the explanatory role for the merely ideal agent that the non-hypothetical bets play for the idealized ideal agent. The worry now is just that there is no good story about how merely hypothetical bets could explain epistemic rationality in the unidealized cases in a practical way, as Gibbard claims they do.

In GRANDMA'S PHONE NUMBER, Sandra knows that knowing phone number won't ever be of practical importance to her. Norms of practical rationality require agents to do the best they can in *actually* satisfying their goals. It hardly seems then that practical rationality directs Sandra to do anything in this case, even if Sandra could merely hypothetically have been faced with the relevant series of bets. More generally, it seems strange to think that when it isn't epistemically possible to the agent that the relevant proposition could play a practical role nonetheless directives of practical rationality apply to the agent in virtue of there being a merely hypothetical series of bets that the agent could face on that proposition. But without the directives of practical rationality in these unidealized ideal cases, Gibbard's cannot explain the agent's epistemic rationality in terms of that agent being practically rational.[14]

Some seem to think that my worry here is mistaken though. In the literature on defenses of probabilist coherence, an appeal to merely hypothetical bets is taken by some to save Dutch Books arguments from a similar worry. There, Dutch Book arguments are given to show that if one's beliefs are probabilistically incoherent, then there is (in a merely hypothetical sense) a series of bets that one would view as fair and guarantee a loss. The objection then is that the merely hypothetical nature of the bets hurts the Dutch Book argument: Just as the merely hypothetical nature

[14] Put another way, the worry here is that merely hypothetical bets can't play the same explanatory role as real bets, since the merely hypothetical ones do not generate practical implications. So then, in trying to have the idealizations satisfy SAVES MAJOR FACTORS, the hypothetical bets in the unidealized cases cannot play the same explanatory role as the bets in the idealized cases.

of the bets above appears problematic for the explanation in Gibbard's unidealized cases, when it comes to Dutch Books, the hypothetical nature of the bets seems to remove the practical implications of the bets from the intended epistemological conclusion.

In response to these worries, several authors have given 'depragmatized' Dutch Book arguments, which (generally speaking) take one's willingness to accept defective bets to expose a kind of defectiveness of the beliefs that would sanction accepting those bets. For these kinds of arguments see Christensen (1996), Howson and Urbach (1993, p. 75–89), and Hellman (1997). One might think that Gibbard could appeal to similar considerations to dispel my worry about his bets having to be merely hypothetical for the unidealized agents.

This seems difficult though. The depragmatized Dutch Books use a hypothetical practical defect to expose, but not explain, a real epistemic defect. Christensen (1996), for example, links beliefs to hypothetical collections of bets via a normative principle about what bets are 'sanctioned' by an agent's credences. Christensen then employs the principle to show that probabilistically incoherent credences sanction defective bets. This then is supposed to expose an inconsistency in the credences themselves. Maher (1997) poses a formal problem for this kind of result, but let's ignore that for a moment. Even if Christensen's argument were right, the result would be that we can expose an *epistemic* inconsistency in the agent's credences by showing that it sanctions a defective collection of bets. Why ought we have probabilistically coherent credences then on this picture? The reason is to avoid *epistemic*, rather than practical, inconsistency.[15] Unlike this kind of approach, Gibbard is attempting to *explain* the epistemic requirements in terms of the practical requirements, not merely use hypothetical practical requirements to expose epistemic requirements that are justified independently. On Gibbard's picture, it is agents "[w]anting truth entirely for the sake of guidance" (2008c, p. 160), a merely hypothetical practical concern for some ideal agents, that is supposed

[15] Christensen makes this quite clear. He says, "Interpreted in this way, Dutch-book arguments do not show that degrees of belief that violate the probability calculus are inconsistent in some previously understood sense. But that is reasonable enough. We need not reduce or assimilate consistency of graded beliefs to some previously understood kind of consistency (such as consistency of all-or-nothing beliefs or of preferences). We are seeking intuitive support for taking a certain set of principles as the best candidate for a formal constraint which plays a role similar to deductive consistency, but which applies to graded beliefs" (p. 457) and later, "Hypothetical vulnerability to guaranteed betting losses is not a practical financial liability. Thus, the force of arguments purporting to derive rules for rational belief from betting-loss considerations must see the hypothetical betting losses as an indicator of a deeper problem. . . . The defect disclosed by Dutch-book vulnerability is then seen as a defect in the beliefs themselves" (p. 478).

to explain their epistemic rationality.

This then is where we stand: The most plausible picture for Gibbard to offer is one in which, for most ideal agents, the agent acting practically rationally with respect to a merely hypothetical series of bets explains the agent being epistemically rational. The worry was that a merely hypothetical practical constraint on an agent don't seem to be able to generate real demands on an agent that may able to explain her epistemic rationality, in the way Gibbard proposes. Similar worries have been posed and responded to for Dutch Book arguments. The responses there though only save the arguments by using the hypothetical practical constraints to expose, rather than justify or explain, the epistemic constraints. Gibbard must claim that the merely hypothetical practical constraints can do more than that. Hypothetical practical constraints must be able to explain or underpin real epistemic constraints. Intuitively, merely hypothetical practical requirements don't seem to be able to *explain* anything as long as they remain merely hypothetical. So, Gibbard needs a story about how this explanation works. How can a merely hypothetical practical constraint on an agent explain, rather than just expose, a real epistemic constraint on that agent?

I worry that the merely hypothetical practical constraints cannot explain real epistemic constraints, as Gibbard would require, but I won't pursue that line any further here. In the next section, I'll undermine Gibbard's motivation for producing this kind of account in the first place by showing that there is alternative way to think about the connection between epistemic rationality and the aim of belief.

4.3 *Aiming at the Truth*

Recall that at the beginning of his argument, Gibbard appeals to what he takes to be a metaphor about belief, that belief aims at the truth. "Belief . . . can't aim literally; it's we who aim," he tells us (2008c, p. 143). Gibbard then uses this idea to motivate his MINIMAL TEST FOR EPISTEMIC RATIONALITY:

MINIMAL TEST FOR EPISTEMIC RATIONALITY "When a person forms her credences with epistemic rationality, is it as if she were choosing her credences voluntarily, rationally aiming, in light of her credences, at truth in those very credences" (2008c, p. 147).

Using this minimal requirement for epistemic rationality, Gibbard then argues that valuing the truth cannot fully explain epistemic rationality, because one can value the truth (satisfy CONDITION T) without thereby having a credence-eliciting utility function, which Gibbard takes to be a requirement that follows from the MINIMAL TEST FOR EPISTEMIC RATIONALITY. Gibbard puts his conclusion like this though:

> Does belief, then, aim at truth? Yes, but in a special way. Belief, we have seen, aims at truth, but not perhaps for the sake of truth itself. Belief aims at truth for the sake of guidance. (2008c, p. 161)

Gibbard concludes that only a species of aiming at the truth could explain epistemic rationality. But strictly speaking, this doesn't follow from what he says in the first part of the argument. There, he only shows that *valuing* the truth, i.e. having a utility function that satisfies CONDITION T, isn't sufficient for epistemic rationality. Gibbard seems to need another premise here, namely that to *aim* at the truth is to *value* the truth. I think Gibbard is intentionally aligning these two notions and that he makes that intention especially clear, for example, when he speaks of utility functions that satisfy CONDITION T (a restriction on what the agent *values*) as "variants of 'aiming at truth"' (2008c, p. 149-50).[16]

In getting his argument started then, Gibbard employs two substantive claims about the popular idiom that belief aims at the truth: First, Gibbard takes the aim to be merely metaphorical – expressing only something about *believers*, not *beliefs*. Second, Gibbard takes it that for an agent to aim at the truth is for her to value accuracy in her credences. This combination of these two claims seems too strong though. If believers aim at the truth and to aim at the truth requires valuing the truth, then agents who deny that they value the truth would seem to fail to have any beliefs. Perhaps such agents as *modelable* as though they value the truth (even in so denying it), but such a model would be deficient at least in being charitable to the agent. So, in this final section, I'll suggest that a externally well-motivated conception of what is required for belief to aim at truth could deny both of Gibbard's claims. I'll also sketch how this kind of account might offer a promising account of how to make sense of epistemic rationality in terms of this newly-conceived-of aim.

4.3.1 Conceiving of the Aim

Gibbard says, "Belief . . . can't aim literally; it's we who aim," and in a sense, he is right. If aiming is an action, no belief can perform it (as only agents can perform

[16] In discussing Gibbard, Swanson (2008, p. 179) seems to follow Gibbard in switching between these two distinct notions. This is particularly clear in Swanson's summary of Gibbard's argument, where Swanson only uses 'value' talk but then puts the conclusion in terms of aims.

actions). But, we can take the idiom to be ascribing a property to belief: "Belief aims at the truth" is short for "belief has an aim of truth." This way of understanding the idiom saves the idea that the idiom is about belief itself, rather than believers, which makes it a more charitable interpretation.[17]

What then does it mean for belief, or any other thing, to have an *aim*? There are a few ways in which something can be said to 'aim at' or 'have an aim of':[18]

- In the most common form of aiming, agents have aims that guide their action. When aims are of the type that can guide our action, they must be conceptually individuated. For example, the moral saint has the aim of promoting the good. She need not also have the aim (in this sense) of producing the greatest amount of utility, even if utilitarianism is true, since judging something to be utility maximizing need not have the same reasons-giving role by the agent's lights as judging it to promote the good.[19] I'll say that these are aims in the *narrow goal* sense.
- Agents also have aims, in the sense of having a goal, but in a less finely-individuated way. For example, suppose Bob is trying to sum 57 and 68. We can say, "he aims to believe that 57 and 68 is 125." We don't thereby attribute to Bob that goal in the narrow sense. Since Bob doesn't know what 57 and 68 is, it doesn't motivate him in the way narrow goals motivate agents.[20] In this *wide goal* sense of 'aim,' an agent has an aim at some goal when the agent is pursing a state of where the goal obtains.[21]
- There is a *teleological* sense in which things have aims. If it is the *telos* or proper function of an object to ϕ, we can describe the object has having an aim of ϕ-ing. A dam in a river, for example, aims at obstructing the flow of water.

[17] Also notice that this way of expressing that some thing has a property is not so uncommon: we might say, for example, "This recipe demands 2 grams of rosewater." In doing so, we're not suggesting that the recipe is actually demanding, like a parent demands a child's obedience; rather, it's the recipe that has a demand.

[18] I am very thankful to the discussion of possible senses of 'aim' by David Plunkett (2012, p. 20-1). Many of these senses of 'aim' are similar to his, but my senses differ from his in some important ways.

[19] For example, we could say to the moral saint, "ϕ-ing produces the greatest amount of utility," to which he could properly respond, "So? I don't aim to produces the greatest amount of utility. I aim to promote the good." So, in this narrow sense of aims, aims are individuated conceptually.

[20] To see that believing that 57 and 68 is 125 is not Bob's *narrow* goal, notice that were he to adopt it in a normal way, he would thereby satisfy it and it would not be his goal. A natural thing to think about this kind of goal, though, is that one has this wider kind of aim at *P* just in case one has an a narrow aim that entails *P*, but I won't take a stand on that here.

[21] There is an intentional sense of 'aim,' in which one aims at some goal when they intend to bring about the goal (Plunkett 2012, p. 20). I believe that the narrow and wide goal senses of 'aim' more finely distinguishes this kind of *intentional* aim.

This is not because the dam adopts any goals or values. Rather, the dam, insomuch as it functions properly, obstructs the flow.

- We also say of objects, such as projectiles, that they are 'aimed' at their targets. An arrow need not have the proper function of hitting the bullseye or be the subject of an agent's intention to hit the bullseye in order to be aimed at the bullseye. The stock market might also be "aiming to make a recovery," in the same sense of 'aim.' In this *trajectory* sense, we say that something has an aim at some object or state of affairs when, under some salient conditions, the object would come in contact with the object or produce the state of affairs.
- Finally, there is a *normative* sense of 'aim.' Aims, in the normative sense, specify standards of success, goodness, or other types of evaluation. Clocks aim at telling the time because a clock is a *good* clock insomuch as it tells the right time. This kind of aim is also sometimes put in terms of correctness conditions: The clock is *correct* when it tells the time. Importantly, the term 'correct' here is distinctly normative; to say that the clock is correct is to say more than it tells the actual time; it also conveys that the clock is as the clock ought to be.[22]

The first two kinds of 'aim,' the narrow and wide goal senses, are kinds that only agents can have, so belief, as opposed to believers, cannot have aims in these senses. Belief also doesn't have an aim of truth in the trajectory sense since a belief cannot literally contact truth. The remaining two senses of 'aim,' the teleological sense and the normative sense, are popularly endorsed in the literature as senses in which belief does have an aim of truth.

The normative conception of the aim of truth is typically formulated as a conceptual truth about BELIEF: it's a conceptual truth that a belief that P is *correct* if and only if P.[23] This claim expresses a standard of correctness for belief. It's important that the sense of 'correct' being used is normative, so that, following Ewing (1939), a correct belief is a belief the one ought to have, for some sense of 'ought.'[24] If belief

[22] We sometimes use the word 'correct' in a non-normative way, as in 'correct supposition.' To call something a 'correct supposition' is just to say that the content of the supposition is true, since suppositions aren't evaluated as good or bad in virtue of whether their contents are true.

[23] For example, see Wedgwood (2002), Shah (2003), Shah and Velleman (2005), Engel (2005), Engel (2008), and Chapter III. In the first paper, Shah only advocates for the 'only if' direction of this claim. In the later paper, the biconditional is advanced. A bit ironically, Gibbard is often taken to espouse this view as well. He says, for example, "For belief, correctness is truth. Correct belief is true belief. . . . Correctness, now, seems normative . . . The correct belief, if all this is right, seems to be the one [a person] ought, in this sense, to have" (2005, p. 338-9)

[24] It's also important to distinguish the normative aim thesis from a similar merely descriptive

aims at the truth in this normative sense, it's a conceptual truth that beliefs ought to be true, for some sense of 'ought.' The sense of 'ought' in play is a distinctly doxastic one, that's part of the very idea of belief. So in this sense, a true belief is a belief how it ought to be, i.e. being true is good *for a belief*.[25]

Proponents of a teleological conception of the aim of belief typically agree with the normative conception that a belief is correct iff its content is true. They disagree with the normative conception in how this is to be understood though. According to the teleological conception of the aim, true belief is correct "*because* only true beliefs achieve the aim involved in believing" (Engel 2012, p. 4). There is value, proponents of the teleological conception argue, in belief functioning properly, i.e. having a true content. This explains why correct belief is true belief, they claim.[26]

Nishi Shah (2003) argues that a purely teleological conception sense of the aim of belief is incompatible with a good explanation of the phenomenon of *transparency*, that "when asking oneself *whether to believe that p*, [one must] immediately recognize that this question is settled by, and only by, answering the question *whether p is true*" (2003, p. 447). Shah (2003) goes on, with Shah and Velleman (2005), to show that the normative conception of the aim of belief can explain the phenomenon. I agree with Shah and Velleman and provide a new argument for the normative conception of the aim in Chapter III. In the next section, I'll suppose that belief does aim in the normative sense, and I'll use this aim to sketch an account of epistemic rationality in terms of this aim.

4.3.2 Saving Epistemic Rationality: A Sketch

Gibbard's argument against understanding epistemic rationality in terms of the aim of belief depends on taking claims about the aim of belief to be metaphorical for claims about what agents value. According to the normative conception of that aim, belief aims at truth in that it's a conceptual truth that belief ought to be

characterization of belief, i.e. that belief is (generally) produced, revised, and destroyed by truth-conducive processes. The normative sense of 'aiming at the truth' is different: it says that beliefs *ought* to be true (independently of whether they are).

[25] Again here, the 'good' is relative to the standards of belief. Similarly, a good clock is one that tells time well. We can evaluate beliefs (and clocks) by other standards, such as how well they promote moral goals or how well they roast chickens, but typically when we say that something is a good clock, we mean that it's good *for a clock*, i.e. relative too the standards against which clocks are evaluated. So, the claim here is the a good belief is on that's true. That's because beliefs are to be evaluated relative to a standard of truth, so a true belief is good *for a belief*.

[26] See Engel (2012) for further discussion of the difference between the teleological and normative conceptions. For defences of the teleological conception, see Velleman (2000, p. 17) (who later gives up the view), Noordhof (2001), and Steglich-Petersen (2006).

true. In contrast to Gibbard, a defender of the normative conception isn't thereby committed there being any constraints on the goals of believers. To be a believer is just to have a belief – an attitude that ought to be true. As far as the aim of belief is concerned, one can be a believer without adopting the goal that one have true beliefs. So, with the normative conception of the aim of belief, we can try anew to make sense of epistemic rationality in terms of the aim of belief.

Ralph Wedgwood takes up this project explicitly and characterizes rationality as a kind of means of achieving the end of having true beliefs (2002, p. 276). He says, "Roughly, rational beliefs are beliefs that either result from, or (in the case of background beliefs) amount to, one's following a rule or set of rules that it is rational for one to believe to be reliable" (2002, p. 282). Norms of rationality, on Wedgwood's view, amount to instruments derivative on the normativity of the aim of belief, since by trying to use reliable processes, we're trying to have true beliefs.

I agree with the general approach of Wedgwood, that of trying to make sense of the normativity of epistemic rationality in terms of the distinctly doxastic normativity of the aim of belief. Here I won't challenge Wedgwood's proposal, but I'll sketch a simpler and more general framework for making sense of epistemic rationality in terms of the aim of belief. My goal is not to give a full defense of this account (or even give all the details); rather, I intend to make the account plausible.

Start with the account of rationality due to Richard Foley (1987). He claims that we can think of different kinds of rationality as rationality with respect to the goals of agents: "rationality is best understood in terms of a person pursuing his goals in a way he would believe to be effective were he to take time to reflect carefully on the question of how best to pursue them" Foley (1987, p. 6). On Foley's view, one is rational in pursuing an end when one is doing the best they can do by their own lights to achieve that goal they have for themselves.

On Foley's account, there are two elements of rationality: there's the end which rationality serves, and there's the conditions under which one counts as rationally pursuing that end (which we may call "the condition of rational pursuit"). According to him, the ends of rationality are goals of the agent, and the condition of rational pursuit is that the agent pursue the goal "in a way he would believe to be effective were he to take time to reflect carefully on the question of how best to pursue them." Both the end of rationality and the condition of pursuit are internal to the agent, in that the end is a goal that the agent adopts for herself and the condition of rational pursuit is dependent on states internal to the agent.

We can take Kelly (2003) as showing that Foley's account of rationality cannot be fully general if it only assess agents relative to their goals.[27] We can generalize Foley's view of rationality to avoid Kelly's concerns by allowing the end of rationality to be any standard, not just a goal of the agent:

λ-RATIONALITY An agent is *rational* with respect to some standard λ just in case they are doing the best they can consistent with their abilities in achieving λ.

On this account of rationality, we can assess the rationality of an agent relative to any standard that the agent might satisfy – independently of whether the agent adopts for herself that standard as a goal. Whereas according to Foley, the ends of rationality are goals of the agent (and are thereby internal to the agent), by treating rationality as merely relative to a standard, we're left with a more externalist understanding of rationality – one that allows us to assess agents as rational or not with respect to any aim, not just the ones the agent sets for herself.[28]

This generalized version of Foley's scheme still admits of the standard cases by substituting in the agent's goals for the standard. For example, an account of practical rationality is produced by letting the standard be the agent's goals:

PRACTICAL RATIONALITY An agent is *practically rational* just in case they are doing the best they can consistent with their abilities in satisfying their goals.

When it comes to generating an account of *epistemic* rationality then, the question is what standards to we hold an agent accountable to when we evaluate them as epistemically rational or not. According to the normative conception of the aim of belief, there is a natural contender for what that standard should be – it's the standard for belief built-in to the very idea of belief. Supposing we can generalize the aim of full belief to the aim of degreed belief or credences, we can say that the aim of degreed belief is to be accurate. The proposal then is that the standard for epistemic rationality is the standard of correctness for (degreed) belief – accuracy.

EPISTEMIC RATIONALITY An agent is *epistemically rational* just in case they are doing the best they can consistent with their abilities in having accurate credences.

[27] Kelly doesn't exactly put his point like this, but this is one lesson of it. Kelly shows that the instrumental conception of epistemic rationality fails, so what follows is that epistemic rationality cannot be made sense of in terms of just the agent's goals. It doesn't follow that a more generalized evaluation of the agent in Foley's style doesn't work.

[28] I suspect that rationality judgments are more external in regard to the condition of rational pursuit than Foley's account suggests. This is why λ-RATIONALITY also generalizes Foley's condition of rational pursuit, making it more external. For the purposes of this sketch, that difference isn't what is doing the work.

So, on this sketch of an account, whether an agent is epistemically rational is a question of whether they are doing the best their abilities will permit them to do in achieving the aim of belief. Being epistemically rational then doesn't require that the agent adopt such a goal for themselves or even intentionally pursue it.

Gibbard's primary focus was on explaining why epistemically rational agents must be as if they have credence-eliciting utility functions. This is because the MINIMAL TEST FOR EPISTEMIC RATIONALITY seems to require that of agents. Being epistemically rational in the sense I propose would require agents to pass this test and be as if they have credence-eliciting utility functions. This is because if the kinds of ideal agents under consideration were doing the best they can to have accurate credences, they would not let themselves be in a position where they take themselves to be able to rationally advance their pursuit of truth by merely adjusting their credences. As I explained in section 4.1.1 above, doing so would not leave the agent in a position to see herself as pursuing the truth to the best of their abilities.

A natural objection to this kind of account might appeal to the fact that epistemic rationality seems to have something to do with one's evidence: a person is epistemically rational, we might say, only if she properly responds to her evidence. The sketch of an account I propose doesn't mention an agent's evidence at all.

My reply to this objection is that the kind of sketch of an account I propose is meant to be a more general understanding of epistemic rationality than the kind that the objector has in mind. Surely, being epistemically rational does require that agents properly respond to their evidence, but this could follow from a more general account. To see that the evidence responsiveness requirement follows from the account, notice that the general account requires that the agent be doing the best she can by her own lights in having accurate credences. An agent's evidence bears on the truth or falsity of something the agent has a doxastic attitude about. So, if they agent is not responding to the evidence, she's not doing the best she can in terms of having accurate credences.

A similar kind of story will work for Wedgwood's account of epistemic rationality summarized above. Wedgwood's account requires that one's beliefs "result from, or (in the case of background beliefs) amount to, one's following a rule or set of rules that it is rational for one to believe to be reliable" (2002, p. 282). Like the evidence responsiveness proposal above, Wedgwood can be seen as offering a precisification of what is required for an agent to count as doing the best they can by their own lights in satisfying the doxastic standard of having true beliefs.

4.3.3 Epistemic Rationality: Saved?

So, can we avoid Gibbard's conclusion that epistemic rationality cannot be made sense of just in terms of the aim of belief by adopting a new conception of that aim? In a sense, we can, but in another sense, we cannot.

As many have recognized, rationality is not a distinctly epistemic notion. We can be rational or not in many different respects – including practically and epistemically. Epistemic rationality is distinguished from other kinds because of its distinctive standard of evaluation. *Epistemic* rationality is rationality evaluated with respect to belief's aim – true beliefs or accurate credences. So, contra Gibbard, in a sense, we can understand *epistemic* rationality in terms of the aim of belief.

Of course, the claim here is that epistemic rationality can be understood in terms of the aim of belief, if we already understand what rationality is more generally. But, epistemic rationality cannot be understood *solely* in terms of the aim of belief. Gibbard's example from the first part of the his argument is enough to show that the standard set by aim of belief is insufficient to generate the norms of epistemic rationality. Epistemic rationality requires agents to be as if they have credence-eliciting utility functions, but all that comes from the aim of belief, on the normative conception, is that accurate credences are correct. Just satisfying this standard for credences doesn't guarantee one a credence-eliciting utility function. So, since being as if one has a credence-eliciting utility function is a requirement of epistemic rationality, that requirement must be explained by what it is for an agent to act epistemically *rationally* – it is the rational pursuit of the epistemic end that explains it, not the aim itself. So, Gibbard is right that the requirements of epistemic rationality cannot be *fully* explained in terms of the aim of belief. Only when the aim of belief functions as a standard for rationality can it explain those requirements.

4.4 Concluding Remarks

In the first part of this chapter, I presented a summary of Gibbard's arguments for two separate conclusions: (1) that epistemic rationality cannot be understood in terms of the aim of belief, and (2) that epistemic rationality is more plausibly explained by the requirement of practical rationality that one maximize prospective guidance value. I then responded to the claims in reverse order.

Once we understand the role of idealizations in Gibbard's proposed explanation, we see that the most plausible version of the claim has the epistemic rationality of some ideal agents being explained in terms of a merely hypothetical practical requirement. It's unclear though how those merely hypothetical practical constraints can *explain* anything about the constraints on the real agent though. Whereas some have taken merely hypothetical practical requirements to expose epistemic requirements, in order for Gibbard to develop his proposal as he has given it, he must go further and show how those merely hypothetical can explain (in a justificatory sense) the epistemic requirements.

Gibbard gets saddled with that explanatory project because in the first part of the argument, he claims that epistemic rationality cannot be understood in terms of the aim of belief. Key to that argument is the idea that 'belief aims at the truth' is metaphorical for a claim about believers' values. In the last section, I suggested that this wasn't the best way to understand to aim of belief. I offered five senses in which something or someone can be said to 'aim.' Two of those senses of aim only apply to agents, but the idiom tells us that it's belief, not believers, that have the aim. So I focused my attention on the remaining three. Of those, two senses of 'aim' are popularly endorsed in the literatures as senses in which belief aims at truth. If we take belief to aim at truth in the normative sense of 'aim,' then I showed that we can begin to see how to make sense of epistemic rationality in terms of that aim. According to the normative reading of the idiom, belief aims at the truth in that true belief is correct belief.

Epistemic rationality, I suggest, is a species of a larger genus, that of rationality. Gibbard is right that not all of the requirements of being epistemically rational are derivable from the aim of belief. Epistemic rationality is the rational advancement towards the standard set by the aim of belief, and what's required to rationally advance toward a standard is more than what is required merely by the standard itself. On the other hand, there is a sense in which we *can* understand epistemic rationality in terms of the aim of belief: What's distinctive about *epistemic* rationality is that it's rationality evaluated relative to the aim of belief.

BIBLIOGRAPHY

Audi, R. (2005). *The Good in the Right: A Theory of Intuition and Intrinsic Value*. Princeton University Press.

Blackburn, S. (1971). Moral Realism. In Casey, J., editor, *Morality and Moral Reasoning*, pages 101–24. London, Methuen.

Boghossian, P. A. (2003). The Normativity of Content. *Philosophical Issues*, 13(1):31–45.

Brandt, R. (1979). *A Theory of the Good and the Right*. Oxford University Press, USA.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press.

Chappell, R. Y. (2005). Inquiry and Deliberation. Blog Post on *Philosophy, et cetera*. http://www.philosophyetc.net/2005/08/inquiry-and-deliberation.html.

Christensen, D. (1996). Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers. *Journal of Philosophy*, 93(9):450–479.

Clifford, W. K. (1877). The Ethics of Belief. *Contemporary Review*.

Davidson, D. (1963). Actions, Reasons, and Causes. *Journal of Philosophy*, 60(23):685–700.

de Finetti, B. (1937). La Prévision: Ses Lois Logiques, ses Sources Subjectives. In *Annales de l'Institut Henri Poincaré 7*, pages 1–68. Paris. Translated into English by Henry E. Kyburg Jr., Foresight: Its Logical Laws, its Subjective Sources. In Henry E. Kyburg Jr. and Howard E. Smokler (1964, Eds.), Studies in Subjective Probability, 53-118, Wiley, New York.

Egan, A. (2007). Quasi-Realism and Fundamental Moral Error. *Australasian Journal of Philosophy*, 85(2):205–219.

Elgin, C. Z. (2005). Review: Williams on Truthfulness. *The Philosophical Quarterly*, 55(219):pp. 343–352.

Engel, P. (2002). *Truth*. Acumen Press.

Engel, P. (2005). Truth and the Aim of Belief. *Laws and Models in Science*, pages 77–97.

Engel, P. (2008). In What Sense Is Knowledge the Norm of Assertion? *Grazer Philosophische Studien*, 77(1):45–59.

Engel, P. (2012). In Defense of Normativism about the Aim of Belief. In Chan, T., editor, *The Aim of Belief*. Oxford University Press.

Enoch, D. (2006). Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action. *Philosophical Review*, 115(2):169–198.

Evans, M. and Shah, N. (2010). Mental Agency and Metaethics. Unpublished Manuscript.

Ewing, A. (1939). A Suggested Non-naturalistic Analysis of Good. *Mind*, 48(189):1–22.

Ewing, A. (1955). Recent Tendencies in Moral Philosophy in Great Britain. *Zeitschrift für philosophische Forschung*, 9(2):337–347.

Firth, R. (1952). Ethical Absolutism and the Ideal Observer. *Philosophy and Phenomenological Research*, 12(3):317–345.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Foley, R. (1987). *The theory of epistemic rationality*. Harvard University Press Cambridge, Mass.

Georgeff, M. P., Pell, B., Pollack, M. E., Tambe, M., and Wooldridge, M. (1999). The Belief-Desire-Intention Model of Agency. In *Proceedings of the 5th International Workshop on Intelligent Agents V, Agent Theories, Architectures, and Languages*, ATAL '98, pages 1–10, London, UK. Springer-Verlag.

Gibbard, A. (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Harvard University Press.

Gibbard, A. (1992). Thick Concepts and Warrant for Feelings. In *Proceedings of the Aristotelian Society*, volume 66.

Gibbard, A. (2002). Normative and recognitional concepts. *Philosophy and Phenomenological Research*, 64(1):151–167.

Gibbard, A. (2003). *Thinking How to Live*. Harvard University Press, Cambridge.

Gibbard, A. (2005). Truth and Correct Belief. *Philosophical Issues*, 15(1):338–350.

Gibbard, A. (2008a). Aiming at Truth Over Time: Reply to Arntzenius and Swanson. *Oxford Studies in Epistemology*, 2.

Gibbard, A. (2008b). Notes on Normativity, Disagreement, and Meaning. Unpublished. Available online at philosophy.fas.nyu.edu/docs/IO/5882/GibbardPaper.pdf.

Gibbard, A. (2008c). Rational Credence and the Value of Truth. *Oxford Studies in Epistemology*, 2.

Gibbard, A. (2012). *Meaning and Normativity*. Oxford University Press. Manuscript.

Goldman, A. (1999). *Knowledge in a Social World*. Oxford University Press.

Goldman, A. (2002). *Pathways to Knowledge: Private and Public*. Oxford University Press.

Goldman, A. (2010). Williamson on Knowledge and Evidence. In Greenough, P., Pritchard, D., and Williamson, T., editors, *Williamson on Knolwedge*. Oxford University Press.

Greco, J. (2003). "Knowledge as Credit for True Belief". In Michael DePaul, L. Z., editor, *Intellectual Virtue: Perspectives From Ethics and Epistemology*, pages 111–134. Oxford: Oxford University Press.

Greco, J. (2009). Knowledge and Success From Ability. *Philosophical Studies*, 142(1).

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and semantics*, volume 3. New York: Academic Press.

Grimm, S. (2009). Epistemic Normativity. In Adrian Haddock, A. M. and Pritchard, D., editors, *Epistemic Value*, pages 243–264. Oxford University Press.

Harman, G. (1977). *The Nature of Morality: An Introduction to Ethics*. Oxford University Press, USA.

Hawthorne, J. and Stanley, J. (2008). Knowledge and Action. *Journal of Philosophy*, 105(10):571–590.

Hellman, G. (1997). Bayes and Beyond. *Philosophy of Science*, 64(2):191–221.

Horwich, P. (1990). *Truth*. Oxford University Press.

Horwich, P. (2000). Norms of Truth and Meaning. *Royal Institute of Philosophy Supplement*, 47(-1):19–34.

Horwich, P. (2006). The Value of Truth. *Noûs*, 40(2):347–360.

Howson, C. and Urbach, P. (1993). *Scientific Reasoning: The Bayesian Approach*. Open Court.

Huemer, M. (2005). *Ethical Intuitionism*. Palgrave Macmillan.

Hume, D. (1978). *A Treatise of Human Nature*. Clarendon Press ; Oxford University Press, Oxford : New York :, 2nd edition.

Jackson, F. (1974). Defining the Autonomy of Ethics. *The Philosophical Review*, 83(1):pp. 88–96.

Jackson, F. (2013). Autonomy of Ethics. In *International Encyclopedia of Ethics*. Blackwell Publishing Ltd.

James, W. (1979). The Will to Believe. In Burkhardt, F., Bowers, F., and Skrupskelis, I., editors, *The Will to Believe and Other Essays in Popular Philosophy*, pages 13–33. Harvard University Press.

Jeffrey, R. (1983). *The Logic of Decision*. University of Chicago Press.

Joyce, J. M. (2009). Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In Huber, F. and Schmidt-Petri, C., editors, *Degrees of Belief*, volume 342, pages 263–297. Springer.

Kelly, T. (2003). Epistemic Rationality as Instrumental Rationality: A Critique. *Philosophy and Phenomenological Research*, 66(3):612–640.

Kolodny, N. and MacFarlane, J. (2010). Ifs and Oughts. *Journal of Philosophy*, 107(3).

Kornblith, H. (1993). Epistemic Normativity. *Synthese*, 94(3).

Korsgaard, C. (2009). *Self-constitution: Agency, Identity, and Integrity*. Oxford University Press.

Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, 1:337–355.

Kratzer, A. (1981). The Notional Category of Modality. In Eikmeyer, H. J. and Rieser, H., editors, *Words, worlds, and contexts. New approaches in word semantics*, pages 38–74. de Gruyter, Berlin.

Kratzer, A. (1991). Modality/Conditionals. In von Stechow, A. and Wunderlich, D., editors, *Semantics: An International Handbook of Contemporary Research*, pages 639–656. de Gruyter, Berlin.

Lewis, D. (1981). Causal Decision Theory. *Australasian Journal of Philosophy*, 59(1):5–30. Reprinted in *Philosophical Papers*, Volume II, pp. 305-337.

Lewis, D. (1996). Elusive Knowledge. *Australasian Journal of Philosophy*, 74(4):549–567.

MacIntyre, A. C. (1981). *After Virtue : A Study in Moral Theory*. University of Notre Dame Press.

Maher, P. (1997). Depragmatized Dutch Book Arguments. *Philosophy of Science*, 64(2):291–305.

McDowell, J. (1981). Non-Cognitivism and Rule-Following. In *Wittgenstein: To Follow A Rule*. Routledge.

Moore, G. E. (1903). *Principia Ethica*. Cambridge University Press, Cambridge.

Nelson, M. T. (1995). Is it always fallacious to derive values from facts? *Argumentation*, 9:553–562.

Noordhof, P. (2001). Believe What You Want. In *Proceedings of the Aristotelian Society*, volume 101, pages 247–265.

Owens, D. J. (2003). Does Belief Have an Aim? *Philosophical Studies*, 115(3):283–305.

Parfit, D. (2011). *On What Matters*. Oxford University Press.

Pigden, C. R. (1989). Logic and the Autonomy of Ethics. *Australasian Journal of Philosophy*, 67:127 – 151.

Plunkett, D. (2012). Legal Positivism and the Moral Aim Thesis. Manuscript Available Online.

Price, H. (1998). Three norms of assertibility, or how the MOA became extinct. *Noûs*, 32(S12):241–254.

Prior, A. N. (1960a). The Autonomy of Ethics. *Australasian Journal of Philosophy*, 38(3):199 – 206.

Prior, A. N. (1960b). The Runabout Inference Ticket. *Analysis*, 21:38–39.

Pritchard, D. (2007). *Epistemic Luck*. Oxford University Press.

Quine, W. V. (1951). Two Dogmas of Empiricism. *Philosophical Review*, 60(1):20–43.

Railton, P. (1997). On the Hypothetical and Non-hypothetical in Reasoning about Belief and Action. *Ethics and Practical Reason*, pages 53–79.

Ramsey, F. P. (1926). Truth and Probability. In Braithwaite, R., editor, *The Foundations of Mathematics and other Logical Essays*. Kegan, Paul, Trench, Trubner & Co., London.

Rawls, J. (1955). Two Concepts of Rules. *Philosophical Review*, 64(1):3–32.

Regan, D. (1980). *Utilitarianism and co-operation*. Clarendon Press.

Russell, G. (2010). In Defence of Hume's Law. In Pigden, C., editor, *Hume on Is and Ought*. Palgrave MacMillan.

Russell, G. and Restall, G. (2010). Barriers to Implication. In Pigden, C., editor, *Hume on Is and Ought*. Palgrave MacMillan.

Schervish, M. (1989). A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879.

Schroeder, M. (2010). Value and the Right Kind of Reason. *Oxford Studies in Metaethics*, 5:25–55.

Schroeder, M. (2011). Attitudes and Epistemics. Unpublished Manuscript.

Shah, N. (2003). How Truth Governs Belief. *Philosophical Review*, 112(4):447–482.

Shah, N. and Velleman, J. D. (2005). Doxastic Deliberation. *The Philosophical Review*, 114(4):497.

Smith, H. M. (2010). Subjective Rightness. *Social Philosophy and Policy*, 27(2):64–110.

Smith, M. (1994). *The Moral Problem*. Blackwell.

Sosa, E. (2007). *A Virtue Epistemology*. Oxford University Press.

Steglich-Petersen, A. (2006). No Norm Needed: On the Aim of Belief. *Philosophical Quarterly*, 56(225):499–516.

Swanson, E. (2008). Note on Gibbard, 'Rational Credence and the Value of Truth'. *Oxford Studies in Epistemology*, 2.

Valentini, L. (2012). Ideal Vs. Non-Ideal Theory: A Conceptual Map. *Philosophy Compass*, 7(9):654–664.

Van Fraassen, B. C. (1980). *The Scientific Image*. Oxford University Press.

Velleman, J. D. (2000). *The Possibility of Practical Reason*. Oxford University Press.

Vranas, P. (2010). Comments on 'Barriers to Implication'. In Pigden, C., editor, *Hume on Is and Ought*. Palgrave MacMillan.

Wedgwood, R. (2002). The Aim of Belief. *Philosophical Perspectives*, 16:267–97.

Wedgwood, R. (2007a). *The Nature of Normativity*. Clarendon University Press.

Wedgwood, R. (2007b). Normativism Defended. In *Contemporary Debates in Philosophy of Mind*. Blackwell.

Weisberg, M. (2007). Three Kinds of Idealization. *Journal of Philosophy*, CIV(12):639–659.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford New York.

Williamson, T. (2003). Blind Reasoning. *Aristotelian Society Supplementary Volume*, 77(1):249–293.

Williamson, T. (2006). Conceptual Truth. *Aristotelian Society Supplementary Volume*, 80(1):1–41.

Williamson, T. (2007). *The Philosophy of Philosophy*. Blackwell Pub. Ltd.

Wright, C. (1994). *Truth and objectivity*. Harvard University Press.

Zagzebski, L. T. (1996). *Virtues of the Mind: An Inquiry Into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge University Press.

Zollman, K. J. S. (2007). The Communication Structure of Epistemic Communities. *Philosophy of Science*, 74(5):574–587.