FALLIBILITY FOR EXPRESSIVISTS

2 · 16 · 19

Abstract

Quasi-realists face the challenge of providing a plausible analysis of acknowledgments of moral fallibility (e.g., *I believe that eating meat is wrong, but I might be mistaken*). According to the standard quasi-realist account, worries about moral fallibility are worries about whether one would retain one's current moral beliefs in idealized conditions. However, this account faces serious difficulties. In light of these difficulties, this paper develops a new expressivist analysis of acknowledgments of moral fallibility, according to which they express moral uncertainty. I show that this 'Credal Analysis' is independently motivated and avoids the difficulties facing the standard account. Finally, I take up the question of how moral expressivists should understand moral uncertainty, developing a view on which an agent's moral credences are understood in terms of the degrees to which they plan to adopt various reactive attitudes.

1 The Challenge

According to moral expressivism, the primary function of moral discourse is not to describe the world. Rather, it's to express desire-like attitudes. The main challenge for expressivism comes from the fact that moral discourse seems to behave much like ordinary, 'descriptive' discourse. Expressivists in the quasi-realist tradition seek to meet this challenge: they aim to reconcile expressivism with the realist trappings of moral language.

Can this ambitious goal be achieved? In recent years, some authors have claimed that quasi-realists face a particularly daunting challenge when it comes to *acknowledgments of moral fallibility*.[1] An example: Ava has just completed a term paper arguing that eating meat is morally wrong. She's confident that her thesis is true. Still, she admits that it's possible that her thesis is mistaken. After all, ethics is hard, and she's aware that many smart people disagree with her. She is thus inclined to say things like:

(1) a. I believe eating meat is wrong.
    b. But I might be mistaken.

How should we understand such acknowledgments, if moral beliefs are just desire-like states? Call this 'The Fallibility Challenge.'

[1] See Egan (2007); Parfit (2011): 395-96; Köhler (2015).

2 The Path Ahead

To their credit, quasi-realists have not left this challenge unanswered. The most well-developed account of acknowledgments of moral fallibility comes from Blackburn. On Blackburn's account, to acknowledge your moral fallibility is to express doubts about whether you would retain your moral beliefs under idealized conditions—that is, conditions of full information, careful reflection, and so forth.[2] However, this 'Idealization Analysis' faces serious difficulties. Egan (2007) and Köhler (2015) argue that it cannot make sense of the full range of acknowledgments of moral fallibility. And Schroeder (2013) objects that it looks *ad hoc* and implausible from a compositional perspective.

In light of these difficulties, this paper develops a new solution to the Fallibility Challenge. I begin with the observation that we typically acknowledge our moral fallibility using epistemic modals (e.g., the term *might* in (1b)). Over the last decade, there has been a surge of interest in expressivist treatments of epistemic modals (Yalcin 2007, 2011; Rothschild 2012; Moss 2013). According to 'credal expressivism,' the primary function of epistemic modals is not to describe the world. Rather, it's to express agents' credences. While credal expressivism and moral expressivism are usually developed in isolation from one another, they make for natural partners. Both are opposed to a descriptivist semantics, and both are motivated on similar grounds. Moreover, I'll argue that by integrating the two, we get a promising new analysis of acknowledgments of moral fallibility. According to what I'll call the 'Credal Analysis,' acknowledgments of moral fallibility are *expressions of moral uncertainty*. I'll try to convince you that this analysis neatly avoids the problems facing the Idealization Analysis: it makes sense of full spectrum of acknowledgments of moral fallibility; it is also compositionally well-motivated.

An obvious objection is that my solution only pushes the problem back a step. My solution requires that expressivists can make sense of moral uncertainty. But can they? In the second half of this paper, I respond to this worry by developing a new expressivist treatment of moral uncertainty. Drawing on ideas from Gibbard (1990, 2003), Sepielli (2012), and Goldstein (2016), I propose that an agent's moral uncertainty consists in the degree to which they plan to adopt various reactive attitudes and actions. If this account is on the right track, we'll have solved two problems for quasi-realists at one fell swoop: we'll have shown how they can make sense of both moral uncertainty and moral fallibility.

3 Idealization and its Discontents

3.1 The Idealization Analysis

Let's begin by taking a closer look at the standard quasi-realist response to the Fallibility Challenge. The central idea is clearly articulated by Blackburn:

[2] See esp. Blackburn (1973, 1998), as well as Horgan and Timmons (2015). See also Ridge (2015) for a variant of this analysis, cashed out in a hybrid expressivist framework.

> How can I make sense of my own fears of fallibility? Well, there are a number of things that I admire: for instance, information, sensitivity, maturity, imagination, coherence. I know that other people show defects in these respects, and that these defects lead to bad opinions. But can I exempt myself from the same possibility? Of course not... So I can think that perhaps some of my opinions are due to defects of information, sensitivity, maturity, and imagination, and coherence. (1998: 318)

Let's see how this applies to our example involving Ava. Expressivists claim that in uttering (1a) (*I believe that eating meat is wrong*) Ava reports a fact about her conative attitudes—for example, that she disapproves of eating meat. The Idealization Analysis maintains that by following up with (1b) (*But I might be mistaken*), Ava acknowledges that this conative attitude might stem from some defect—a failure of information, imagination, etc. An improved sensibility—a sensibility that is better informed, more imaginative—might not share this conative attitude.

3.2 First Difficulty: Idealized Sensibilities Might Err

Despite its appeal, the Idealization Analysis faces two serious difficulties. First, as Egan (2007) forcefully argues, it has difficulty accommodating the thought that even *idealized* sensibilities might err. To unpack this worry, imagine that Ava ruminates as follows:

> It seems an agent could know all of the non-moral facts, while still failing to know all of the moral facts. Of course, this failure could be due to further defects on the agent's part—say, lack of imagination or sensitivity. But need it be? This isn't obvious: just because an agent is fully informed, sensitive, and imaginative, it's not clear that she'll be guaranteed to know all of the moral truths.

Following this line of thought, Ava judges:

(2) Even if my belief that eating meat is wrong survives idealization, this belief might be mistaken.

The Idealization Analysis predicts that (2) expresses an incoherent judgment: it amounts to judging that even if her belief survives idealization, it might not survive idealization. This is an uncomfortable result. On the face it, Ava's concern seems perfectly coherent.[3]

[3] My statement of the worry departs in a number of respects from Egan's formulation, in part so as to sidestep certain features of Egan's formulation that Blackburn (2009) deems objectionable. In particular, Blackburn insists that the notions of 'idealization' and 'improvement' are themselves normative notions, and hence should be understood in expressivist fashion. My formulation of the challenge is perfectly consistent with an expressivist construal of such notions.

Proponents of the Idealization Analysis might push back on this: maybe once we get clear on the relevant sense of 'idealization,' we'll recognize that (2) is incoherent after all. Blackburn takes this line in reply to Egan. According to Blackburn, if a belief is false, '[T]hen an improvement is clearly on the cards, namely, replacing it with the truth' (2009: 206). And so any belief that survives idealization is, by definition, guaranteed to be true.

But this response comes with a serious cost. If quasi-realists pursue this line, they forfeit any claim to have provided a non-circular analysis of acknowledgments of moral fallibility. On the view that results, an acknowledgment that a moral judgment $\mathcal{J}$ might be in error is analyzed as an acknowledgment that $\mathcal{J}$ might not survive idealization. But 'idealization' is itself analyzed in terms of the avoidance of moral error.[4]

3.3 Second Difficulty: Semantic Plausibility

Even if Egan's objection can be overcome, trouble is still in store. Note that Ava's utterance of (1) is an instance of a more general schema:

(3) a. I believe ϕ.
    b. But I might be mistaken.

While the metaethics literature has focused on moral instances of this schema, it's easy to come up with non-moral instances. Suppose Ava and her friends are planning to see a movie. Ava seems to recall reading that the movie starts at 7, but she isn't positive. We can imagine her saying:

(4) a. I believe the movie starts at 7.
    b. But I might be mistaken.

This suggests a general constraint on a solution to the Fallbility Problem. Any explanation of the coherence of (1) should generalize to all instances of (3), including (4). The natural way to do so is to give a compositional semantic analysis of the possibility modal *might*, and to derive one's analysis of (1) from this more general analysis.

But the Idealization Analysis does not pursue this strategy. Rather than giving a general semantics for possibility modals, it proceeds in a 'piecemeal, *ad hoc* fashion' (Schroeder 2013: 416). Consequently, it generates implausible predictions concerning how embedding a sentence under *might* affects its meaning. As Schroeder notes, the Idealization Analysis holds that the following two sentences are equivalent:

(5) It might not be wrong to eat meat.

[4]Some may suggest that expressivists shouldn't try to offer a non-circular analysis of acknowledgments of moral fallibility. (See Blackburn (2009) for a suggestion in this vein.) However, as Köhler (2015) and Ridge (2015) emphasize, one goal of the quasi-realist program is to explain the mental states we express when we make various realist-sounding claims. This commits quasi-realists to giving some story about what mental state Ava expresses when she utters (1b). Without some analysis of acknowledgments of moral fallibility, it's unclear how the quasi-realist can fulfill this commitment.

(6) I might not disapprove of eating meat if I were an idealized agent.

However, expressivists typically take pains to insist that the following sentences are *not* equivalent:

(7) It is not wrong to eat meat.

(8) I would not disapprove of eating meat if I were an idealized agent.

After all, expressivists maintain that there is a crucial difference between expressing a mental state *m* and reporting that one is in *m*. According to the expressivist, (7) expresses lack of disapproval of eating meat; it does not report that one would not disapprove of eating meat under such-and-such conditions. But then why would embedding (7) and (8) under the same epistemic modal (*might*) generate equivalent sentences?[5]

4 The Credal Analysis

In order to develop a more satisfactory solution, I propose proceeding in two steps. The first is to heed Schroeder's injunction: rather than treating acknowledgments of moral fallibility in isolation, let us give a general account of the semantic contribution that *might* makes to any linguistic environment in which it occurs. Once we have an analysis of *might*, we can then leverage this into a general analysis of acknowledgments of fallibility—both moral and non-moral.

In order to implement this strategy, I'll start by discussing a particular analysis of *might* that has attracted considerable attention in recent years: credal expressivism. I'll show that credal expressivism fits very naturally with moral expressivism. And it has the resources to solve the Fallibility Challenge.

4.1 Credal Expressivism

Credal expressivism is typically advanced in opposition to a more traditional 'descriptivist' semantics for epistemic modals, according to which utterances of epistemic modals purport to describe facts about the world. Consider:

(9) It might be raining.

According to descriptivists, (9) describes the world as being a particular way—in particular, as being one where the prejacent (*It's raining*) is consistent with the information available to some contextually determined agents (for example, the speaker).[6]

[5] Schroeder's objection assumes that (6) is the result of embedding (8) under *might*. And this in turn appears to assume that the occurrence of *might* in (6) takes wide scope over the whole conditional. But this is open to doubt. Arguably, the logical form of (6) is *If ϕ, then might ψ*, rather than *Might (If ϕ, then ψ)*. However, this worry about Schroeder's argument only exposes a more basic problem with the semantic plausibility of the Idealization Analysis. After all, the question as to the logical form of (6) only arises because (6) is a conditional. But (5) is not. Why then would we expect (5) and (6) to be equivalent?

[6] See e.g., Kratzer (1981); Dowell (2011).

Credal expressivist deny this. According to credal expressivists, (9) does not have descriptive truth conditions at all. Instead, it expresses the speaker's positive credence in the possibility that it's raining.

To my knowledge, moral and credal expressivism have never been explicitly integrated—at least not in a way that allows epistemic modals to embed moral vocabulary.[7] However, moral and credal expressivism make for natural allies. For starters, the two views are structurally analogous. Both claim that certain bits of language are not used to describe the world, but rather to express certain mental states. Furthermore, both views are typically motivated by similar arguments.

First, there's what I'll call 'the subject matter argument.' Credal expressivists often object that descriptivism offers an implausible picture of the subject matter of modal beliefs. Having a modal belief, according to this objection, does not require having beliefs *about* anyone's body of information. Yalcin gives the example of Fido, who believes his owner might give him a bone (2011: 308). Does this require that Fido believes that the receipt of a bone is compatible with his information (or anyone else's, for that matter)? This seems like a stretch. More plausibly, it only requires Fido assigning some credence to the possibility that his owner will give him a bone.[8]

In a similar vein, moral expressivists often object that rival accounts provide an implausible picture of the subject matter of moral beliefs. Take, for example, a simple subjectivist or relativist view, according to which S believes that stealing is wrong iff S believes that S disapproves of stealing. Against this, expressivists often complain that the *reasons* for believing that stealing is wrong are importantly different from the reasons for believing that one disapproves of stealing. The latter sort of reasons are psychological in nature—they are the sort of reasons that could be provided by a psychoanalyst or a brain-scan. By contrast, the former sort of reasons are moral: they are reasons to disapprove of stealing in the first place.

Second, there's what I'll call 'the argument from disagreement.' Building on Price (1983), Yalcin (2011) argues that if I disagree with your claim that it might be raining, I do not thereby disagree with the claim that *your* epistemic state is compatible with the possibility that it will rain. Here too, parallel arguments have been offered for moral expressivism. Against simple subjectivism, it is often objected that if I disagree with your

[7] In his 2007 paper, Yalcin mentions Gibbard (1990) as a point of inspiration; however, the framework developed there does not synthesize moral and credal expressivism. In other work, Yalcin develops a semantic implementation of plan-expressivism about deontic modals (2012). However, he does not integrate the semantics for deontic modals with his semantics for epistemic modals in a way that makes sense of sentences containing both expressions (e.g., *It's possible that we ought to donate more to charity*).

Recently, Ridge (2015) has integrated moral expressivism with an account of epistemic modals that bears some resemblance to credal expressivism. However, there is an important difference. For Ridge, a sentence containing an epistemic modal doesn't express an agent's credences; rather, it expresses a normative perspective, and, on top of that, makes a representational claim about which credences would, given a potential body of evidence, be permitted by acceptable normative standards. This raises a number of questions—for example, how children who lack the concepts of credences and standards could form modal beliefs—which do not arise for credal expressivism.

[8] See also Rothschild (2012); Moss (2013).

claim that stealing is wrong, I do not thereby disagree with the claim that *you* disapprove of stealing.[9]

The upshot: those who find the arguments for moral expressivism convincing should also be drawn to credal expressivism.

4.2 Solving the Fallibility Challenge

On to the second step of the solution. Credal expressivism delivers a general analysis of acknowledgments of fallibility. Take any instance of (3). According to what I'll call the 'Credal Analysis,' in asserting (3a) (*I believe* ϕ), the speaker asserts that she believes ϕ. But in following up with (3b) (*But I might be mistaken*), she hedges by expressing her positive credence in $\neg\phi$, and hence her lack of certainty in ϕ.

I propose this holds for both moral and non-moral instances of ϕ. Thus when Ava utters (1a), she starts by reporting her belief that eating meat is wrong. And in following up with (1b), she hedges this by expressing her lack of certainty that eating meat is wrong.

Of course, this proposal immediately raises the question: 'How should expressivists understand this sort of uncertainty?' I will devote the next two sections to answering this question. But let us first see how the Credal Analysis avoids the difficulties facing the Idealization Analysis.

The first problem for the Idealization Analysis was that one can coherently worry that idealized sensibilities might err, as expressed in Ava's utterance of (2) (*Even if my belief that eating meat is wrong survives idealization, this belief might be mistaken*). The Credal Analysis has no trouble here. According to the Credal Analysis, (2) expresses Ava's conditional uncertainty. Specifically, it expresses the state of assigning some positive credence to the prospect that eating meat isn't wrong, conditional on her belief that eating meat is wrong surviving idealization. Since this is a coherent credal state, the Credal Analysis correctly predicts the coherence of (2).

Some might wonder whether the problem re-emerges in a different form. Does the Credal Analysis predict that it would be incoherent for Ava to worry that her moral *certainties* might be mistaken? If so, isn't this prediction equally problematic?

In assessing this objection, we need to take care: much depends on how we formulate the relevant worry. One way is *via* a conditional such as:

(10) Even if I were certain that eating meat is wrong, I might still be mistaken.

The Credal Analysis can make perfect sense of this worry. (10) expresses Ava's positive credence that eating meat isn't wrong, conditional on her counterfactual self becoming certain that eating meat is wrong. And this is surely a coherent credal state: one can rationally assign some credence to the prospect of becoming certain of a falsehood.

Alternatively, Ava could worry that one of her current certainties is mistaken:

(11) It's possible that I'm certain of something which is nonetheless mistaken.

[9] See e.g., Stevenson (1937); Schroeder (2010): 69-70.

According to the Credal Analysis, (11) expresses the state of having some credence that one is currently certain of a falsehood. Plausibly, certainties are not luminous: one can be rationally uncertain as to what one believes with certainty. And so (11) also expresses a coherent credal state.

A final way of trying to express the relevant worry is to affirm both that one is certain of some claim and that this claim might be mistaken:

(12) a. I'm certain that eating meat is wrong.
b. ? But I might be mistaken.

However, *this* discourse sounds odd—worse, at any rate, than (10) or (11). The Credal Analysis explains why. If (12a) were true—if Ava were (absolutely) certain that eating meat is wrong—then she could not coherently assign some credence to the possibility that eating meat is not wrong. And so she could not coherently be in the mental state that is expressed by (12b).[10]

On to the second problem for the Idealization Analysis: it looks implausible and *ad hoc* from a compositional perspective. Here too the Credal Analysis is in the clear. Take any possibility modal, combine it with some sentence ϕ, and you'll get a sentence that expresses the speaker's positive credence in ϕ. Consequently the Credal Analysis avoids predicting that (5) (*It might not be wrong to eat meat*) is equivalent to (6) (*I might not disapprove of eating meat if I were an idealized agent*). Indeed, (5) is not equivalent to any claim about what agents might disapprove of under such-and-such conditions.[11]

5 Credences for Expressivists

5.1 Smith's Challenge

At this point some readers may be growing impatient. 'Clearly,' readers may observe, 'The Credal Analysis will only solve the problem if expressivists can make sense of moral credence. But can they do so? If moral beliefs are just desire-like states, then what is it to have, say, a .8 credence that eating meat is wrong?'

[10] However, she could coherently assign a high credence to (12a) while being in the mental state expressed by (12b). This would occur if Ava falsely believed that she were certain that eating meat is wrong.

[11] Some may wonder whether we couldn't reap the same benefits without going expressivist. Why not say that an utterance of *Might* ϕ describes the world as one where the speaker assigns some positive credence to ϕ? I think that this 'descriptivist Credal Analysis' would indeed avoid the two main difficulties with the Idealization Analysis. All the same, it makes for a much more awkward partner to moral expressivism, for the reasons given in §4.1. It runs afoul of the subject matter argument, since it makes modal beliefs *about* our own mental states. Likewise with the disagreement argument: I can disagree with your claim that it might be raining without thereby disagreeing with the claim that you assign some credence to the possibility that it is raining. Thus anyone who takes these arguments to cut against moral descriptivism should also take them to cut against a descriptivist Credal Analysis. For further difficulties, see the embedding data in Yalcin (2007).

A preliminary observation: quasi-realists were already committed to providing an account of moral credence, independent of the Fallibility Challenge. After all, quasi-realists aim to make sense of ordinary moral discourse. And we often say things like:

(13) Darren is confident that lying is wrong, but he's even more confident that stealing is wrong.

So if quasi-realism has any hope of success, we had better be able to give some account of moral credence. Thus proponents of the Credal Analysis do not incur any *new* explanatory debts by appealing to moral credences in order to solve the Fallibility Challenge.

Dialectical points aside, what should expressivists say about moral credence? An initially tempting strategy is to identify moral belief with some conative state that comes in degrees, such as desire or disapproval. Expressivists could then hold that moral credences correspond to degrees of this state. However, Smith (2002) shows that any such approach would conflate two distinct dimensions of moral judgment. The first—'certitude'—is your degree of confidence that, say, eating meat is wrong. The second—'importance'—is *how* wrong you take eating meat to be. These dimensions can vary independently: you could have a middling credence that eating meat is very wrong, or a high credence that it is is slightly wrong. Any account that reduces moral credence to degrees of desire or disapproval will lack the structural resources to draw this distinction.

Smith's argument poses a formidable challenge to expressivist accounts of moral uncertainty. But I think it would be premature to give up hope. In what follows, I'll sketch what I think is the most promising solution.

5.2 The Reactive Attitudes Account

In order to introduce my account, it will be helpful to make a brief foray into an independent problem facing expressivists: the 'Moral Attitude Problem' (Miller 2003). This problem arises from the fact that not every desire makes for a *moral* belief. I can desire a cookie without thinking it would be morally wrong for me to go cookie-free. The challenge for expressivists, then, is to say what sort of desire-like state makes for a distinctly moral judgment.

One promising answer is that moral judgments—or, at least, wrongness judgments—are distinguished by their connection with the reactive attitudes, such as guilt and outrage, as well as what we might call the 'reactive actions,' such as praise and punishment. This idea has its roots in Mill (1861). According to Mill, to claim that so-and-so acted wrongly is to say that he ought to be punished, 'if not by law, by the opinion of his fellow-creatures; if not by opinion, by the reproaches of his own conscience.'[12] Gibbard (1990) gives Mill's idea an expressivist twist. On Gibbard's view, to judge that eating meat is wrong is to adopt some approval-like attitude towards a syndrome of reactive attitudes and actions

[12]Mill (1861): chp. V, par. 4.

towards meat-eating—a syndrome that might include resentment and outrage towards meat eaters, as well as admiration for vegetarians.

This 'Reactive Attitudes Account' offers a straightforward solution to the Moral Attitude Problem. Why doesn't my desire for a cookie constitute a wrongness judgment? Because I do not approve of blaming or shaming myself in the event that I fail to consume one.

But the real payoff of the Reactive Attitudes Account is that it answers Smith's challenge, as demonstrated by Sepielli (2012). According to the Reactive Attitudes Account, wrongness judgments involve two components: (i) a cluster of reactive attitudes and actions, (ii) a conative attitude towards this cluster. As Sepielli notes, we can use these two components to model Smith's two dimensions of moral judgments. Variations in importance consist in variations in the severity of the reactive attitudes/actions. Variations in certitude consist in variations in the strength of the conative attitude towards these reactive attitudes/actions.

To illustrate, consider a simple (too simple, for reasons to be discussed shortly) version of the Reactive Attitudes Account, according to which the relevant reactive attitude/action is blame, and the relevant conative attitude is *pro tanto* approval.[13] Then having a high credence that eating meat is slightly wrong amounts to having a high degree of *pro tanto* approval for mildly blaming meat eaters. And having a middling credence that eating meat is very wrong amounts to having a moderate degree of *pro tanto* approval for severely blaming meat eaters.

While the Reactive Attitudes Account is not the only possible expressivist model of moral credence,[14] I think it has considerable appeal. Not only does it answer Smith's challenge, it is also independently motivated *via* consideration of the Moral Attitude Problem.[15]

Does this mean our work is done? Not quite. A further challenge remains…

6 Coherence Constraints on Moral Credences

6.1 A Residual Challenge

As Bykvist and Olson (2009, 2012) point out, moral certitude has a lot in common with descriptive certitude. For example, both have upper and lower bounds, corresponding to complete certainty and complete disbelief. Not all conative attitudes are like this. Take

[13]The *pro tanto* qualification is needed since one can judge S acted wrongly without judging that S is blameworthy. On the present framework, this would be captured by saying one does not *ultima tanto* approve of blaming S, since one's *pro tanto* approval of blaming S is overridden by one's awareness of various exculpatory factors. For further discussion, see Gibbard (1990): 43-47.

[14]For alternative accounts, see Eriksson and Olinder (2016); Ridge (forthcoming).

[15]Yet more independent motivation comes from Frege-Geach style worries. See Schroeder (2008): chp.4 for an argument that the Reactive Attitudes Account helps expressivists handle negation. (For critical discussion, see Baker (2018).)

desire. Arguably, there is no upper bound to the degree to which one could desire some outcome.

This is just the tip of the iceberg. The more general problem is that moral and descriptive certitude seem to be subject to probabilistic coherence constraints:

Normalization A rational agent's credence in a tautology should be 1.
Non-Negativity A rational agent's credence in any proposition should be ≥ 0.
Additivity A rational agent's credence in a disjunction of mutually exclusive propositions should be the sum of their credences in each of the disjuncts.

Plausibly, these constraints do not only apply to descriptive credences; they also govern moral credences. But this is surprising if moral and descriptive credences are very different sorts of states.

As Staffel (forthcoming) stresses, expressivists owe us an explanation for this striking coincidence. A natural strategy is to develop a more detailed account of the conative attitude that constitutes moral certitude—an account that explains why moral and descriptive certitude both have probabilistic structure.

6.2 Moral Certitude as Degrees of Planning

Here I'll sketch one candidate account. As before, I take my cue from Gibbard. In his recent work, Gibbard suggests that moral belief is a type of plan (2003). Combined with the Reactive Attitudes Account, this leads to the following view. To believe that eating meat is wrong is to have a *pro tanto* plan that calls for adopting various reactive attitudes and actions towards eating meat.[16]

One advantage of identifying moral beliefs with plans—rather than, say, desires—is that plans seem to share many important features with descriptive beliefs. Both involve a phenomenology of 'settling' some question. If Ava believes the movie starts at 7, then she has settled for herself the question of whether it starts at 7. Similarly, Bratman observes that if I plan to spend all afternoon at the library, I take myself to have settled the question of where to spend my afternoon (1987: 18-19). And plans, much like descriptive beliefs, can agree or disagree with one another. To illustrate with one of Gibbard's examples, suppose Holmes plans to pack. Suppose also that Moriarty's plan, contingent on being in Holmes' situation, is to refrain from packing. Then, Gibbard contends, there is a natural sense in which their plans disagree.

Do plans also resemble descriptive beliefs in their coherence requirements? This seems plausible when it comes to outright plans. Just as it is incoherent to have inconsistent

[16]Some might object to the idea that we can *plan* to adopt reactive attitudes. Aren't the reactive attitudes emotional states, and don't these lie outside of our voluntary control? However, we *do* sometimes intend to regulate our emotions: we might form a New Year's resolution to relinquish a grudge, or remain calm in the face of adversity. Moreover, even if we grant the objection, we could tweak the account to put all the weight on reactive *actions*—e.g., criticism, social sanctions, and the like. Surely there is no obstacle to talking about plans to engage in such actions.

outright beliefs, so it is incoherent to have inconsistent outright plans: if I plan to spend the afternoon at the library, it would be incoherent to also plan to stay home all day. In what follows, I'll draw on ideas from Goldstein (2016) to argue that a similar parallel holds between degrees of belief and degrees of planning.

First things first: do plans even come in degrees? To motivate this, suppose Jim plans to do two things on Monday: submit his grades and finish a referee report. However, Jim realizes he might not be able to do both. Failure to submit his grades will have dire consequences; failure to finish the referee report will not. Thus he plans to submit his grades to a greater degree than he plans to finish the referee report. He is fully committed to the former; he is less than fully committed to the latter.

As Goldstein (2016) argues, reflection on the nature of plans provides further support for the idea that plans come in degrees. The most detailed account of plans in the current literature is due to Bratman (1987). On Bratman's view, planning is a multi-track dispositional state: intending to ϕ involves being disposed to ϕ, being disposed to avoid reconsidering whether to ϕ, and being disposed to seek out means of ϕ-ing. Suppose we accept this conception of plans, at least as a working hypothesis. Now, most dispositions are gradable. A statue can be more or less fragile; a material can be more or less flammable. The dispositions that constitute planning are no exception. If Jim plans to submit his grades to a greater degree than he plans to finish his referee report, we'd expect him to be more strongly disposed to submit his grades than to finish the report. We'd also expect him to be more strongly disposed to avoid reconsidering whether to submit the grades than to avoid reconsidering whether to finish the report. Following Goldstein, we can suggest that the degree to which an agent plans to ϕ is determined by the strengths of the various dispositions that constitute planning to ϕ.

Putting all of this together: let ϕ be the property of adopting some particular cluster of reactive attitudes and actions towards meat eating, for example, *mildly blaming meat eaters*. According to the current proposal, an agent has credence 1 that eating meat is slightly wrong iff they fully plan to ϕ. And this obtains iff they have all of the dispositions that constitute planning to ϕ to the strongest degree possible. On a Bratmanian view, this will involve being fully disposed to mildly blame meat eaters, being fully disposed to avoid reconsidering whether to mildly blame meat eaters, and being full disposed to seek out the means of mildly blaming meat eaters. An agent has credence 0 that eating meat is slightly wrong iff they fully plan to $\neg\phi$. And this obtains iff they have all of the dispositions that constitute planning to $\neg\phi$ to the strongest degree possible, for example, being fully disposed to refrain from mildly blaming meat eaters, being fully disposed to avoid reconsidering whether to refrain from mildly blame meat eaters, etc.

The next step is to appeal to an independent line of argument, developed in Goldstein (2016), that a dispositional view of plans leads naturally to *probabilism about plans*: the idea that degrees of planning are subject to probabilistic constraints. While the details are complex, here is a broad-brushstrokes overview of the argument. In the literature on dispositions, one promising approach is to analyze dispositions in terms of quantification over some domain of cases C in which some stimulus condition obtains (Manley and

Wasserman 2008; Vetter 2014). A major selling point of these 'modal accounts' is that they shed light on the gradability of dispositions. On a modal account, the degree to which x has some disposition D is determined by the proportion of C-cases in which x manifests D. For example, suppose we want to analyze the degree to which a statue is fragile. The C-cases will be circumstances in which the statue is dropped from various heights, or struck with varying degrees of force. The higher the proportion of such cases in which the statue breaks, the more fragile the vase.

Combining a dispositional account of plans with a modal account of dispositions allows us to explain why moral certitude has upper and lower bounds. The upper bound for one's credence that eating meat is slightly wrong corresponds to manifesting whatever dispositions constitute planning to ϕ in 100% of some domain of cases C. The lower bound corresponds to manifesting whatever dispositions constitute planning to $\neg\phi$ in 100% of C-cases. As a corollary, it follows that moral certitude satisfies the Non-Negativity axiom. Since one cannot manifest the dispositions that constitute planning to $\neg\phi$ in more than 100% of C-cases, one's credence that eating meat is slightly wrong cannot be less than 0.

This approach also explains why moral certitude obeys Normalization. On our view, for Ava to have credence 1 in the tautology, *Either eating meat is wrong or it isn't*, is for her to fully plan to bring about the tautology: *Either I blame meat eaters or I don't* ($\top$). And this amounts to manifesting the dispositions that constitute planning to satisfy $\top$ in all C-cases. On a Bratmanian view of plans, the relevant dispositions will be: (i) a disposition to satisfy $\top$, (ii) a disposition to search for the means necessary to do so, (iii) a disposition to avoid reconsidering whether to do so. As Goldstein observes, a case can be made for thinking every rational agent will manifest these three dispositions in every C-case, for any plausible candidate for 'C'. After all, it is metaphysically necessary that every agent satisfies $\top$. Thus, every agent will trivially have the means to satisfy $\top$; moreover, it would be pointless for an agent to reconsider whether to satisfy $\top$.

Finally, an analogous argument shows why moral certitude obeys Additivity. According to the present proposal, Ava's credence in the disjunction, *Either eating meat is very wrong or it is not at all wrong*, is the degree to which she plans to bring about the disjunction, *Either I severely blame meat eaters or I don't blame them at all* ($\vee$). And this in turn is reflected in proportion of C-cases in which she manifests the dispositions that constitute this plan. Goldstein argues, again using a Bratmanian view of plans, that the proportion of C-cases in which a rational agent manifests these dispositions will be the sum of the cases in which she manifests the dispositions that constitute planning $\vee$'s disjuncts.

Taking stock: combining dispositionalism about plans with a modal account of dispositions provides one way of motivating probablism about plans.[17] And probabilism about plans allows us to answer the residual challenge facing the Reactive Attitudes Account.

[17] Goldstein also provides an independent argument that degrees of plans are subject to probabilistic requirements. This argument involves extending accuracy dominance arguments for probabilism to degrees of intentions. In the interest of space, I defer discussion of this argument to another occasion. For related discussion, see Staffel (forthcoming).

We will have given a substantive account of the conative state that constitutes moral certitude, an account that explains why it is subject to the same coherence requirements as descriptive certitude.

7 Remaining Questions

This paper has presented a new solution to the Fallibility Challenge: the Credal Analysis. While I hope to have convinced you that the Credal Analysis shows promise, I recognize that it raises a number of further questions. I'll conclude by briefly considering two of the most pressing.

7.1 Conditional Moral Uncertainty

First, some might wonder whether we've fully explained what sort of mental state someone is in when they worry that idealized sensibilities might err. According to our analysis, when Ava asserts (2) she expresses a state of conditional moral uncertainty. While we went on to give a story about moral uncertainty, we have not yet explained *conditional* moral uncertainty.

However, there is a natural way of bridging the gap. We often form conditional plans (or 'contingency plans,' in Gibbard's parlance). Currently I do not plan to buy groceries, since I don't know whether the store is open. But, conditional on the store being open, I plan to buy them. Conditional plans also come in degrees: conditional on the store being open, I plan to buy avocados to a greater degree than I plan to buy caviar. Just as moral credences are identified with degrees of plans, we can identify conditional moral credences with degrees of conditional plans. Thus when Ava asserts (2), she is expressing a sort of *conditional indecision*. Even conditional on her belief that eating meat is wrong surviving idealization, she is less than fully committed to blaming meat eaters.

Some might worry about the plausibility of this diagnosis. Isn't this an odd state of mind for someone to be in? In ordinary cases of conditional indecision, there is typically some further information that would help us settle on a course of action. By contrast, Ava is undecided about whether to blame meat eaters even conditional on knowing all the potentially relevant information. Ava is thus in a state of *fundamental conditional indecision*—a sort of indecision that, by her own lights, no further information could resolve.

But I think expressivists can dispel the apparent oddity of this state by pointing to other examples of fundamental conditional indecision. Suppose Ava is trying to decide on a career: surgeon or philosopher? Perhaps Ava could make up her mind if she learned further descriptive facts (for example, facts about salaries or work-life balance). But need her indecision be so straightforwardly resolved? On another way of fleshing out the case, even if she were fully informed about the two career paths, she would still feel torn. On this way of fleshing out the case, Ava is in a state of fundamental conditional

indecision. I propose that it is this state of mind that expressivists should use when trying to understand Ava's worry that an idealized sensibility might err.

7.2 The Formal Semantics

A second question is how to formally implement my analysis. I've motivated the Credal Analysis *via* compositional considerations. However, I have not actually given a compositional semantics for *might* and *wrong* that delivers the Credal Analysis as a consequence. Can this be done?

I think the answer is *yes*. As a proof of concept, let me briefly sketch an integration of a plan-expressivist semantics (à la Gibbard 2003) with an information-sensitive semantics for epistemic modals (à la Yalcin 2007).

On Gibbard's framework, semantic contents are sets of *world, hyperplan* pairs, where a hyperplan is a formal device that models the content of a maximally decided planning state (2003: 54). To leverage this into a compositional semantics for *wrong*, assume that some hyperplans specify plans to blame various actions to various degrees. Taking our cue from the standard analysis of gradable adjectives in the semantics literature, we could propose that *wrong* denotes a function from actions to degrees of blame.[18] And so the semantic content of a sentence such as:

(14) Eating meat is wrong.

is the set of w, h pairs such that h includes a plan to blame for eating meat to a degree that meets some threshold t (which may be determined by context). Call this set 'WRONG':

$$\text{WRONG} = \{\langle w, h\rangle \mid h \text{ plans to blame for eating meat at } w \text{ to a degree } \geq t\}.^{19}$$

The next step is to integrate this analysis with a compositional semantics for *might*. Here I'll take my cue from Yalcin (2007). Whereas Gibbard enriched ordinary possible worlds propositions with hyperplans, Yalcin enriches them with *information states*. An information state is a set of points representing live possibilities. Simplifying slightly, we can think of these as the possibilities to which the relevant agent assigns positive credence. Yalcin takes these points to be worlds. But since we are interested in modeling moral uncertainty, we will take these points to be world, hyperplan pairs. Following Yalcin, we can hold that *might* existentially quantifies over these points.

To illustrate, consider (5) (*It might not be wrong to eat meat*). The semantic content of this sentence is a set of world, hyperplan, information state triples. Specifically, it is the set of such triples where the information state includes at least one live possibility

[18] The standard analysis of gradable adjectives in the semantics literature analyzes them as functions from entities to degrees on some associated scale (Kennedy 2007). For example, *expensive* denotes a function from entities to units of cost.

[19] One nice feature of this analysis is that it allows us to straightforwardly represent differences in importance. Differences in importance correspond to differences in the threshold for degrees of blame.

(itself a world, hyperplan pair) that does not include a plan to blame for eating meat (to a degree $\geq t$):

$$\Diamond\neg\textsc{wrong} = \{\langle w, h, s\rangle \mid \exists\langle w', h'\rangle \in s\colon \langle w', h'\rangle \notin \textsc{wrong}\}.$$

As Yalcin notes, we can think of this content as representing a property of an agent's credal state. It is the property of assigning positive credence to a world, hyperplan pair where the hyperplan doesn't include a plan to blame for eating meat. And this, in turn, can be thought of as the property of being less than fully committed to blaming for eating meat. According to the present proposal, this is the state of mind that Ava expresses when she utters (5).

This is surely not the only way of formally implementing the Credal Analysis. But it should suffice to show that there are no principled obstacles to developing a compositional semantics that integrates moral and credal expressivism.

References

Derek Baker. Expression and Guidance in Schroeder's Expressivist Semantics. *Erkenntnis*, 83: 829–852, 2018.

Simon Blackburn. Moral Realism. In Casey, editor, *Morality and Moral Reasoning*. Methuen, London, 1973.

Simon Blackburn. *Ruling Passions*. Clarendon Press, Oxford, 1998.

Simon Blackburn. Truth and *A Priori* Possibility: Egan's Charge Against Quasi-Realism. *Australasian Journal of Philosophy*, 87(2):201–213, 2009.

Michael Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.

Krister Bykvist and Jonas Olson. Expressivism and Moral Certitude. *Philosophical Quarterly*, 59: 202–215, 2009.

Krister Bykvist and Jonas Olson. Against the *Being For* Account of Normative Certitude. *Journal of Ethics and Social Philosophy*, 6(2):1–8, 2012.

J.L. Dowell. A Flexible Contextualist Account of Epistemic Modals. *Philosophers' Imprint*, 11(14): 1–25, 2011.

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

John Eriksson and Ragnar Francén Olinder. Non-Cognitivism and the Classification Account of Moral Uncertainty. *Australasian Journal of Philosophy*, 94(4):719–735, 2016.

Allan Gibbard. *Wise Choices, Apt Feelings*. Harvard University Press, Cambridge, MA, 1990.

Allan Gibbard. *Thinking How to Live*. Harvard University Press, Cambridge, MA, 2003.

Simon Goldstein. A Preface Paradox for Intention. *Philosophers' Imprint*, 16(14):1–20, 2016.

Terry Horgan and Mark Timmons. Modest Quasi-Realism and the Problem of Deep Moral Error. In Johnson and Smith, editors, *Passions and Projections: Themes from the Philosophy of Simon Blackburn*. Oxford University Press, Oxford, 2015.

Chris Kennedy. The Semantics of Relative and Absolute Gradable Adjectives. *Linguistics and Philosophy*, 30:1–45, 2007.

Sebastian Köhler. What is the Problem with Fundamental Moral Error? *Australasian Journal of Philosophy*, 93(1):161–165, 2015.

Angelika Kratzer. The Notional Category of Modality. In Eikmeyer and Rieser, editors, *Words, Worlds, and Contexts: New Approaches in Word Semantics*. W. de Gruyter, Berlin, 1981.

David Manley and Ryan Wasserman. On Linking Dispositions and Conditionals. *Mind*, 117(465): 59–84, 2008.

J.S. Mill. *Utilitarianism*. 1861.

Alexander Miller. *Introduction to Contemporary Metaethics*. Polity, Cambridge, 2003.

Sarah Moss. Epistemology Formalized. *Philosophical Review*, 122(1):1–43, 2013.

Derek Parfit. *On What Matters*, volume 2. Oxford University Press, Oxford, 2011.

Huw Price. Does 'Probably' Modify Sense? *Australasian Journal of Philosophy*, 61(4):396–408, 1983.

Michael Ridge. I Might be Fundamentally Mistaken. *Journal of Ethics and Social Philosophy*, 9(3), 2015.

Michael Ridge. Normative Certitude for Expressivists. *Synthese*, forthcoming.

Daniel Rothschild. Expressing Credences. *Proceedings of the Aristotelian Society*, 112(1):99–114, 2012.

Mark Schroeder. *Being For: Evaluating the Semantic Program of Expressivism*. Oxford University Press, Oxford, 2008.

Mark Schroeder. *Noncognitivism in Ethics*. Routledge, New York, 2010.

Mark Schroeder. Two Roles for Propositions. *Noûs*, 47(3):409–430, 2013.

Andrew Sepielli. Normative Uncertainty for Non-Cognitivists. *Philosophical Studies*, 160:191–207, 2012.

Michael Smith. Evaluation, Uncertainty, and Motivation. *Ethical Theory and Moral Practice*, 5(3): 305–320, 2002.

Julia Staffel. Expressivism, Normative Uncertainty, and Arguments for Probabilism. In Szabó Gendler and Hawthorne, editors, *Oxford Studies in Epistemology*, volume 6. Oxford University Press, forthcoming.

Charles L. Stevenson. The Emotive Meaning of Ethical Terms. *Mind*, 46:14–31, 1937.

Barbara Vetter. Dispositions without Conditionals. *Mind*, 123(489):129–156, 2014.

Seth Yalcin. Epistemic Modals. *Mind*, 116(464):983–1026, 2007.

Seth Yalcin. Nonfactualism about Epistemic Modality. In Egan and Weatherson, editors, *Epistemic Modality*. Oxford University Press, Oxford, 2011.

Seth Yalcin. Bayesian Expressivism. *Proceedings of the Aristotelian Society*, 112(2):123–160, 2012.