Truth and Fallibility in Morality: On a New Way Forward for Moral Expressivists

ABSTRACT: Moral expressivists say that 'Eating meat is wrong' expresses a motivational state, like disapproval of eating meat, rather than a representational state, like the belief that eating meat is wrong. How, then, can the expressivist make sense of the possibility that her moral judgements could be mistaken – if they do not aim to represent, how can they misrepresent? This is the challenge from moral fallibility. I am concerned to defend Simon Blackburn's account of fallibility in terms of *potential for improvement*, especially from Andy Egan's charge that it commits the expressivist to an implausible asymmetry between herself and others, or "smugness". I argue that while we ought to concede that the moral truths are *epistemically constrained* – that is, in principle accessible to a suitably placed inquirer – a "suitably placed" inquirer is *anyone* who occupies a node within a vast network of moral outlooks. In making this case, I establish necessary and sufficient conditions on moral truth by expressivist lights; which in turn shows how the expressivist is entitled to a substantive theory of moral *truth*, which I suggest can serve as the explanatory basis for a truth-conditional semantics for moral discourse.

I. Introduction

Some of the sounds we make – groans, cries, sighs, hums, coughs, cheers, tuts – are not meaningful, in the way the words and sentences we use are meaningful. They can, of course, like any other behaviour, impart information or be used to successfully communicate something, but they do not do so by virtue of what they mean. Words and sentences have *semantic properties*, and it is the job of a semantic theory to tell us what these semantic properties are. Words might *refer* to particular entities, like the Queen of England, or *ascribe* particular properties, like baldness. Sentences might have *truth conditions*. But it is not an intrinsic fact about the noises someone makes when they utter 'The Queen of England is bald' that they refer to the Queen of England, ascribe the property of baldness, or are true just in case a specific person has insufficient hair. That *this* noise has semantic properties at all, and that it has the specific semantic properties that it has, cries out for explanation. This is the job of *metasemantics*. (And the distinction applies not just to the noises we make, but to inscriptions and gestures too.)

Focus now on *moral* discourse in particular: discourse about what is morally right or wrong, morally good or bad, or what we ought, morally speaking, to do. Why is the vegetarian's utterance of 'Eating meat is wrong' meaningful? An answer to the general metasemantic question here – the question of how this discourse gets to be meaningful *at all* – that is paradigmatically *expressivist* is to say that moral language is used, as a matter of convention, to express moral attitudes, and that these attitudes are motivational, desire-like, states. The predicate 'is wrong' might be used to express *disapproval*, for example, while 'is right' expresses *approval*. The vegetarian's utterance thus expresses her disapproval of eating meat.

Such a view, I submit, faces two core sets of challenges. The first is to execute the particular metasemantic project given the general starting point: how are we to explain the semantic behaviour of moral sentences – why they have the semantic properties they have – given that they express desire-like states? The second concerns the objectivity or mind-independence of morality. How do we make sense of morality as something that places binding constraints on our behaviour if all we are doing in moral discourse is giving voice to which things we happen to approve or disapprove of?

This paper is primarily concerned with the latter challenge in what I take to be its most incisive form: the question of how the expressivist is to make sense of the possibility that her own moral views may be mistaken. This is the challenge from moral fallibility. I am concerned to defend Simon Blackburn's account of fallibility in terms of *potential for improvement*, especially from Andy Egan's (2007) charge that it commits the expressivist to an implausible asymmetry between herself and others, or "smugness". Roughly, the charge is that we end up thinking of our own epistemic position as privileged, and quite arbitrarily. I argue that while we ought to concede that the moral truths are *epistemically constrained* – that is, in principle accessible to a suitably placed inquirer – a "suitably placed" inquirer is *anyone* who occupies a node within a vast network of moral outlooks. To pin a charge of smugness onto the expressivist, it turns out, requires taking on a number of contentious commitments that, I argue, raise uncomfortable questions for everyone.

But I also want to suggest that the right view on this matter offers a new way for the expressivist to proceed with her metasemantic program. In particular, in arguing that the moral truths are epistemically constrained, I establish necessary and sufficient conditions on moral truth by expressivist lights. We can thus see how the expressivist is entitled to a substantive (though "anti-realist") theory of moral *truth*, which can serve as the explanatory basis for a truth-conditional semantics of moral discourse. In the next section (II), I run through the metasemantic challenge and the proposed way forward in more detail. The following sections (III-X) then turn to addressing the challenge from fallibility.

II. Expressivism and Metasemantics

That expressivism might be understood as specifically a metasemantic thesis has only recently been foregrounded in the literature.[1] By contrast, it has sometimes been thought of as a semantic thesis, as a theory of what the meanings of moral sentences *are*; and in particular as a competitor to a truth-conditional theory of meaning.[2] Such a view inevitably faces those challenges that march under the broad banner of the “Frege-Geach Problem”, which includes the challenge of providing a compositional but non-truth-conditional semantics for moral (and mixed) discourse. One advantage of the metasemantic construal is thus that it is by itself semantically neutral: it doesn’t say anything about what the meanings of moral sentences are, and it therefore doesn’t rule out any particular semantics, including a truth-conditional one. And if the semantics of moral discourse *is* truth-conditional, then the Frege-Geach Problem in the form just mentioned doesn’t arise.[3] Such compatibility is to this extent a desideratum. Nonetheless, what is far from obvious is how the metasemantic expressivist might be *entitled to* a truth-conditional semantics, given her starting point.

Compare the expressivist’s position with that of a representationalist who explains the meaning of ‘Eating meat is wrong’ in terms of a (full-blown) *belief* that it expresses: the belief

[1] By e.g. Chrisman (2012) and Ridge (2014). That it is *best* understood as such is not something I will argue here, though the argument of this paper may well contribute to such an argument.

[2] See e.g. Schroeder (2008, 2010).

[3] More precisely, providing a compositional semantics for moral discourse is no more complicated for the expressivist than it is for anyone else who wishes to use a truth-conditional semantics. It is the elimination of this asymmetry that constitutes dialectical progress.

that eating meat is wrong. On this account, the sentence can inherit its truth conditions from the propositional content of the belief it expresses.[4] This option is unavailable or unattractive for the expressivist. Even if disapproval is a propositional attitude, its content is not the content that we want 'Eating meat is wrong' to have.[5]

If the propositional content of the mental state it expresses is not to serve as the explanatory basis for the truth-conditional content of a moral sentence, then what? How are we to proceed with the metasemantic project? Must we deny that the right semantics is truth-conditional, and face the Frege-Geach Problem head on?

It is here that I want to propose a new way forward for the expressivist. The proposal is to use a substantive notion of moral *truth* as the explanatory basis. Suppose we have one: moral truth consists in some property *F*. We then have a straightforward route to explaining the truth conditions of atomic moral sentences: 'p' is true just in case 'p' is *F*; we can assign propositional content to 'p' by checking at which worlds 'p' is F. The semantics for logically complex sentences can then be understood compositionally, in the standard way.

That a substantive theory of moral truth can play this theoretical role is clear. But what is not clear is that the expressivist is entitled to any such theory. Indeed, some expressivists may balk at the suggestion that they endorse a substantive theory of moral

[4] Lewis (1975) is a good example of someone who endorses this broadly "head-first" approach to explaining the nature of representation, i.e. explaining linguistic representation in terms of mental representation. We are then owed an account of mental representation, but this is everyone's challenge. (Lewis, e.g., uses rationalising interpretation for this next step.)

[5] Disapproval is just a stand-in mental state, but I trust it is clear that this point endures whichever desire-like state the expressivist endorses.

truth, whatever its nature. One popular, "quasi-realist" variety of expressivism, for instance, is at least caricatured as conjoining a *deflationary* construal of truth (and facthood, and miscellaneous other realist-sounding notions) with an expressivist explanation of moral discourse, in order to try and explain the apparently realist surface-form of moral discourse. I think expressivists are better off not going down this road.[6] But rather than making that negative case, I'm going to pursue the positive one, by arguing directly that expressivists are entitled to a substantive theory of moral truth.

My strategy is as follows. I start with the challenge from fallibility, which every expressivist must face, and Blackburn's response. As we chase down the implications of Blackburn's theory, I argue, we find ourselves committed to necessary and sufficient conditions on moral truth: for any atomic moral 'p', 'p' is true just in case the state it expresses

[6] I'm yet to find any good reason for expressivists to be tempted by deflationism. First, for a perspicuous articulation of why mere *expressivism-plus-deflationism* is insufficient qua explanation of moral discourse, see Dreier (1996). Second, while I can see the case for a *minimal* construal of the relevant apparently realist *concepts* (that is, the concept TRUTH, the concept FACT, and so on) – to provide a neat explanation of why we don't hesitate to use this terminology within the discourse in question – minimalist explanations of concepts are quite compatible with the substantivity of the corresponding *properties*. (Consider e.g. Wright (1992) on truth.) Third, what's important to expressivism *qua* anti-realism, it seems to me, is that moral truths and facts are not afforded particular explanatory roles (in explaining our capacity for moral judgement, for instance), not that there aren't any – or any substantive – truths or facts. I thus struggle to see any peculiarly expressivist motivation for taking on the deflationist's (in my view implausible) metaphysics. Nonetheless, *certain* substantive theories seem to be off the table, at least at this stage; e.g., correspondence theories, which would appeal to moral facts or properties as a "worldly" relatum. Given this, one goal of this paper is to show that the expressivist is entitled to the particular substantive theory I defend, by showing that it follows from her metasemantic starting point, given the other commitments defended here.

is *weakly super-stable*. Given this, there is no significant additional theoretical cost attached to endorsing weak super-stability as an account of the nature of moral truth. Expressivists who understand fallibility in the broadly Blackburnian way defended are thus *already* entitled to the requisite theoretical machinery. The point of the preceding is that there is a significant theoretical *gain* to this machinery: an explanatory basis for a truth-conditional semantics for moral discourse. These theses together thus point to a promising new way forward for the moral expressivist.

III. The Challenge from Fallibility

At first pass, there can seem something obnoxiously parochial about the expressivist's theory. In uttering 'Eating meat is wrong', is the vegetarian really only expressing that she happens to disapprove of eating meat? Doesn't this in some way make morality merely a matter of preference? But attempts to make this intuitive worry more precise are often uncompelling. After all, the vegetarian is only "only" expressing that she "happens" to disapprove of eating meat in the same sense that in saying 'Iron is magnetic' I am "only" expressing that I "happen" to believe that iron is magnetic. This does not make the magnetism of iron a subjective or mind-dependent matter, or in any way a mere matter of belief. Schroeder (2014) calls this the "parity thesis": the relation of *expressing* that holds between 'Eating meat is wrong' and disapproval is the same relation that the representationalist says holds between 'Iron is magnetic' and the belief that iron is magnetic.

The fact that in engaging in moral *discourse* we are expressing our own mental states is plainly insufficient to render *morality* subjective or mind-dependent in any problematic sense.[7]

Within this family of worries, the challenge from fallibility strikes me as by far the most penetrating. That we are fallible within the moral domain is a given. Fallibility consists in (i) the possibility of being *mistaken*: that there is some *p* such that you think that *p* when it is not the case that *p*; and (ii) the possibility of being *ignorant*: that there is some *p* such that you do not think that *p* when *p*. The expressivist can make sense of *third-personal* fallibility up to a point simply by making sense of disagreement. To think that *you're* mistaken is just to think that *p* when you think that *not-p*; this might be to, say, disapprove of eating meat where you tolerate it, for instance.[8]

The primary challenge arises from *first-personal* fallibility. The question is how the expressivist can make sense of the possibility that she is, herself, at this very moment, either mistaken or ignorant. If the functional role of moral judgements is motivational rather than representational, then how can moral judgements *mis*represent? We can even play this game on the expressivist's home territory by asking after the nature of the fallibility *judgements*.

(1) Eating meat might not be wrong.

[7] See e.g. Blackburn (2010), Schroeder (2014), and Köhler (2014) and the citations therein. As these sophisticated discussions make clear, there is much more subtlety to the debate concerning mind-dependence or subjectivism than I convey here; but my concern is really with the challenge from fallibility.

[8] Schroeder (2008) maintains that the only way to make sense of inconsistency between states is in terms of two states of the same kind with inconsistent content; but Baker & Woods (2015) argue on the contrary that inconsistency between different kinds of state is a perfectly familiar phenomenon.

Even the most committed vegetarian, when being humble and acknowledging her own fallibility, may assert (1). But what state does she thereby express? What is it to *think* that one of your own moral judgements may be in the wrong?[9]

She is not expressing any uncertainty as to whether or not she actually disapproves of eating meat. She might be quite confident of that, and in any case it is not *that* that she is worried about. Similarly, she cannot merely be admitting the possibility that she might change her mind. By her current lights, in doing so she would be moving from a true judgement to a false one. If someone disagrees with you you're committed to *them* being in the wrong; why should a future time-slice of yourself be any different to someone who disagrees with you right now? Moreover, it may be precisely when you realise that you are *un*likely to ever change your mind (perhaps realising how stubborn you can be) that you become *most* worried about your potential fallibility.

Could she be expressing that she is not *certain* that eating meat is wrong? In a sense, this is exactly what she is expressing; but, unfortunately, unsupplemented this constitutes no progress. For variable confidence is a *belief-like* feature. The expressivist is thus not immediately entitled to it, but must rather explain it. Desire-like states too come in variable degrees: you can disapprove of something *more* than you disapprove of something else. But, as Michael Smith (2002) has argued, this more readily maps onto a distinct feature of moral judgement, which he calls "importance". I might be *equally confident* that, say, lying to my loved ones and lying to my co-workers is wrong, but *also* think that lying to my loved ones is

[9] One strand in Blackburn's (2009) response to Egan (2007) is to complain that Egan focuses on moral error itself, rather than moral judgements. Köhler (2015) responds that Egan's challenge recurs in this setting. I bypass this epicycle by focusing on judgements throughout. While I'm sympathetic with Köhler's conclusion, I think there is much more to Blackburn's response than just this.

worse. The natural (and, I think, correct) thing to say here is that I disapprove of lying to my loved ones *more*; but then variation in strength of desire is explaining variation in judgements of import, not variation in certainty or confidence. The expressivist thus owes us an explanation of what it is to have variable confidence in one's moral judgements before this move is of any use.[10]

If the expressivist cannot make room for first-personal fallibility, the theory is to be rejected.[11] Blackburn addresses the challenge in the following way:

> 'The problem comes with thinking of myself (or of us or our tradition) that I may be mistaken. How can I make sense of my own fallibility? Well, there are a number of things I admire: for instance, information, sensitivity, maturity, imagination,

[10] The flipside of this worry is that the account of fallibility discussed here "doubles-up" as an account of variable confidence; I resist setting this out for reasons of space. Note that Ridge's (2015) alternative to Blackburn's account of fallibility relies on an antecedent entitlement to variable confidence. It is an advantage of the present account that it does not do so, though Ridge's proposal may have other strengths. See also fn.16. In endorsing such an explanation of variable confidence, the expressivist ought to reject the reductivist thesis whereby to think that p *is* to have sufficiently high confidence that p, for moral judgements. Where variation in confidence is explained in terms of a higher-order attitude, the moral judgement *itself* is a motivational, desire-like state. (Thus the view does not collapse into a "hybrid", expressivism on this front.) She can, however, endorse intuitive norms on the relation between the digital and analogue attitudes, e.g., that one ought to think that p iff one has sufficiently high confidence that p.

[11] That this is problematic is obvious, though the objection itself could, I suppose, be developed in diverse ways. One might press that an explanation of moral discourse that cannot make sense of fallibility just looks like a bad explanation. Or that absence of first-personal fallibility is implausibly hubristic. Or that *no one* being able to make sense of their own fallibility perhaps implies a kind of wild relativism. Etc.

> 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 (that would be unpardonably smug). So I can think that perhaps some of my opinions are due to defects of information, sensitivity, maturity, imagination, coherence. If I really set out to investigate whether this is true, I stand on one part of the (Neurath) boat and inspect the other parts.' (1998: 318; see also Blackburn 1993: 20-2)

It is not merely that the vegetarian may change her mind, but that in doing so she may be *improving* her moral outlook, with respect to those epistemic standards (information,[12] sensitivity, etc.) that she values; and which she values precisely because shortcomings with respect to these standards in others leads to bad moral outlooks (outlooks that she disagrees with).[13] It is important that the vegetarian uses *her own* standards here. We are not playing 'the fake externalist game of trying to certify our values without using values' (Blackburn 1996: 89), by trying to understand fallibility by appeal to something (like a realm of moral facts) outside our own values. The game is being played internally.

A couple of clarifications. While we shouldn't expect which standards we endorse to be immediately transparent to introspection, I take it that it is extremely plausible that

[12] In order to avoid begging the question by appealing to moral truths that potentially diverge from our own moral beliefs, 'information' must here be understood as information pertaining to *non-moral* matters.

[13] Compare third-personal fallibility: we want it to be possible that, when we disagree with someone, sometimes we are in the right, and sometimes they are in the right. The natural way of understanding the latter is in terms of our interlocutor's moral outlook being *better* (or agreeing with one that is better) than our own.

ordinary moral agents do endorse epistemic standards in this sense. We do, after all, criticise ourselves and others for basing moral judgements on misinformation, or for being insufficiently sensitive or coherent. We can likewise expect that agents may endorse "meta-standards" for weighing improvements by the lights of different standards against each other. Moreover, note that endorsing a standard like coherence does not require that one know exactly what coherence consists in. This matters, since it highlights that one may be *mistaken* about whether or not a moral outlook is an improvement by the lights of the standards you endorse. A moral outlook might strike us, at least initially, as coherent, when in fact on closer inspection it turns out that it is not so. While the agent's subjective *endorsement* of a standard is crucial in the account of fallibility, whether or not a moral outlook is *actually* an improvement according to your standards can nonetheless be an entirely objective matter. It is *actual* improvement by the lights of the standards she endorses that the vegetarian is worried about.

It will be useful to semi-formalise this idea. Label our agent's judgement set 'B_a'. Let 'b_v' label the state that is expressed by 'Eating meat is wrong' – disapproval of eating meat, say. So $b_v \in B_a$. And let 'S_a' label the epistemic standards that our agent endorses. According to Blackburn, when our vegetarian asserts (1), she is expressing uncertainty as to whether or not there is some set B, such that B is an improvement on B_a by the lights of S_a (which we will label '$B >_{Sa} B_a$') and $b_v \notin B$.

Strictly, this only allows us to understand judgements concerning what I have called *being mistaken*, when fallibility also incorporates *being ignorant*. But the account has a natural extension. Suppose our agent is not a vegetarian, but asserts (2):

(2) Eating meat might be wrong.

On our account, (2) expresses uncertainty as to whether or not there is some $B >_{S_a} B_a$ such that $b_v \in B$.

IV. Egan's Challenge from Smugness

Andy Egan (2007) has raised a worry for this strategy, which draws on the role that our agent's standards – that is, the standards she happens to endorse – play in this account. According to Egan, this gives rise to an implausible asymmetry between the account's first- and third-personal implications. The argument is as follows.

First, consider b_v. Since our agent is a vegetarian, $b_v \in B_a$. But now suppose that there is no $B >_{S_a} B_a$ such that $b_v \notin B$; that is, there is no change that the agent can go through that is an improvement by the lights of the standards she endorses that displaces that belief.[14] Egan calls such a belief, a *stable* belief. On the present account of fallibility, it seems, our agent then cannot make sense of the possibility that her belief is mistaken. Of course, she cannot know *a priori* that any of her moral beliefs *are* stable in this way; but she can know *a priori*, it seems, that if any of her beliefs are stable, then they cannot be in error.

Now take a further agent, with belief set B_b, who disagrees with our agent. Let $\neg b_v$ be the state expressed by 'Eating meat is not wrong', so $\neg b_v \in B_b$. This other agent will also

[14] For ease, I'm going to start talking about moral judgements as "beliefs", but given the expressivist framework so-called "moral beliefs" should strictly be understood as motivational states. (It is not uncommon for expressivists to allow that the states expressed by moral sentences are beliefs in some, typically minimalist, sense – see e.g. Horgan & Timmons (2006).)

endorse some epistemic standards, S_b. But now suppose that there is no belief set $B >_{Sb} B_b$ such that $\neg b_v \notin B_b$; so this rival belief is stable for this agent. This agent thus cannot make sense of the possibility that his rival belief is in error. But given that the two agents disagree, and assuming that where there is (moral) disagreement there is error, at least one of them must be wrong.[15] Since neither agent can make sense of the possibility that their own belief is mistaken, each is committed to the other's belief being so. This puts the expressivist in an uncomfortable position. It seems that none of us can make sense of the possibility that a belief is stable *by our own lights* and yet in error; but we can make sense of the possibility that a belief is stable *by someone else's lights* and yet in error. We are then each committed, by our own lights, to being immune to a kind of error that everyone else is susceptible to. That looks, in Blackburn's words, "unpardonably smug".

Note, however, that the mere *formal* possibility of this asymmetry is insufficient to establish a problem. Suppose that any stable belief of yours is guaranteed to be a stable belief of mine, and *vice versa*. Then the asymmetry between us is eliminated. But the expressivist would have quite some work to do to establish that there is any such general guarantee. After all, as Egan points out, our agents might have radically different standards and/or radically different starting beliefs. We will return to this avenue of response in due course.

Now, a version of Egan's worry does arise within Blackburn's framework, but not precisely the one Egan presents. For in his response to Egan, Blackburn (2009) shows how the expressivist *can* make sense of the possibility that a stable belief of their own is in error. Recall that the agent's standards, S_a, appear in the account of fallibility because our agent

[15] While my own inclinations are absolutist, I do not here want to rule out combining expressivism with relativism. I will thus offer a relativist spin on the discussion in several places below.

endorses them. As mentioned, whether or not a belief set is better or worse by the lights of S_a may be a matter independent of whether or not our agent endorses the standards in S_a. But the view that a belief set that is better by the lights of S_a is an *improvement* is a *value* judgement. And it is a value judgement that the agent can be wrong about.

How does the expressivist make sense of this potential for being mistaken to endorse a standard? Blackburn's suggestion is that we do so in terms of potential for improvement by the lights of the *other* standards the agent endorses. Take some standard $s \in S_a$, and let b_s be the agent's endorsement of this standard; so $b_s \in B_a$. Then let S_a* be the other standards in S_a apart from s ($S_a^* = \{s_x \mid s_x \in S_a \ \& \ s_x \neq s\}$). The fallibility judgement is then the judgement that there may be some $B >_{S_a*} B_a$ such that $b_s \notin B$. That is, that there may be some change that is an improvement by the lights of the other standards you endorse where you stop endorsing that standard. Since this goes for *any* of your standards, you can make sense of the possibility that you are mistaken to endorse any particular standard. Indeed, one may hereby make sense of the possibility that *all* of your standards are mistaken: there may be a further improvement that replaces one standard; and another improvement that replaces another standard; and another that replaces another; and so on. The same thing goes for all of our beliefs. Note that we make sense of this possibility of radical, global error in terms of *incremental* replacement through improvement, rather than a wholesale replacement; hence, the Neurath boat metaphor is especially apt here. (*How* the expressivist can make sense of the sceptical scenario will be important later.)

This allows us to make sense of the possibility that one of our beliefs is stable by the lights of the standards we endorse, and yet in error. For even if one of our agent's beliefs b_v is stable by the lights of the standards in S_a, there may be some *improved* set of standards,

S_a+, such that b_v is not stable relative to S_a+ – i.e., there is some B $>_{S_a+}$ B_a such that $b_v \notin B$.[16] (Note that we do not make sense of this possibility by sudden appeal to some set of standards that is the *correct* one to endorse, as seen from some external, God's-eye perspective. The game is still being played internally.)[17]

However, now we can reintroduce Egan's worry.[18] For what if b_v *is* still stable relative to any improved set of standards, S_a+? Well, our agent cannot rule out that it may yet be unstable relative to some improved version of that improved set, S_a++. But what if b_v is *still* stable relative to any such S_a++? The dialectic iterates. So now suppose that b_v is stable relative to S_a, S_a+, S_a++, S_a+++, and so on; that is, it is stable relative to every improvement on S_a, and every improvement on them, and every improvement on *them*, and so on. Call such a belief, *super-stable*. While our expressivist cannot know *a priori* that any belief of their own

[16] Since acknowledging the fallibility of our normative judgements regarding standards amounts to being less than completely certain that they are correct (fn.10), avoiding the stability limitation does, in this sense, make use of variable confidence in some normative judgements. However, it does not (unlike e.g. Ridge 2015) assume a *distinct* such account. Thanks to an anonymous referee for suggesting this clarification.

[17] Note that our agent is now potentially using changes that are not S_a-approved, but only S_a+-approved, to understand her own fallibility, and such changes are *not* presently endorsed by our agent as improvements. But there is no reason to think that our agent can *only* use changes she presently endorses as improvements to understand her own fallibility. Indeed, the lesson from Egan's stability limitation worry is that this notion is potentially *too* constrained. As long as we are not invoking something like moral facts to make sense of improvement and hence fallibility, we are not playing the representationalist's "fake externalist game".

[18] The following way of reintroducing Egan's worry was first noticed, to my knowledge, in Daniel Elstein's doctoral thesis, and the term 'super-stable' is his – see Elstein (2013: §3.1).

is super-stable, it seems that they cannot make sense of any moral belief of their own being super-stable and yet in error.

Once again, we ought not to forget about ignorance, for a parallel phenomenon arises here. Suppose that our agent doesn't hold some moral belief b: $b \notin B_a$. She can make sense of the possibility that she is ignorant because there may be some $B >_{Sa} B_a$ such that $b \in B$. But suppose there is not. The agent can still make sense of the possibility that she is ignorant, because there may be some improved version of S_a, S_a+, such that there is some $B >_{Sa+} B_a$, and $b \in B$. But suppose that this too is not so. And likewise for any further improvement on S_a. Call such a belief, an *inaccessible* belief. Just as our agent cannot make sense of the possibility that their belief is super-stable and yet in error, so they cannot make sense of the possibility that a belief might be inaccessible to them and yet correct.[19] Call these together, *the Super-Stability Limitation*.[20]

Moving forwards, it will be useful to have the following terminology. Let a *self-endorsed improvement* be a change from a belief set B to a belief set B', such that B' is an improvement on B by the lights of the standards endorsed in B, where this includes changes in the standards endorsed in B that are improvements by the lights of the other standards

[19] In his response to Egan, Blackburn (2009: 206) suggests the following standard for improvement, which could potentially avoid these limitations if the expressivist were entitled to it: 'if [any belief] were false, then an improvement is clearly on the cards, namely replacing it with the truth'. But the very question is how the expressivist can make sense of the possibility that a belief is not true when it is (super-)stable. This standard presupposes the very gap between (super-)stability and truth that we cannot make sense of.

[20] It's worth emphasising that, if your belief b is super-stable, the Super-Stability Limitation does *not* mean that you will not or cannot stop having that belief. It is just that no such change is possible *through self-endorsed improvement*. But you might hit your head, or forget, or just go through some bad reasoning.

endorsed in B. And for any two belief sets, B and B’, we will say that B’ > B just in case there is some *series* of self-endorsed improvements from B to B’.[21] A belief b is thus super-stable at a belief set B just in case b ∈ B and there is no B’ > B such that b ∉ B’; a belief b is inaccessible at a belief set B just in case b ∉ B and there is no B’ > B such that b ∈ B’. If B_a is our agent’s belief set, for some moral b ∈ B_a, a fallibility judgement is the judgement that there may be some B > B_a such that b ∉ B; for some moral b ∉ B_a an ignorance judgement is the judgement that there may be some B > B_a such that b ∈ B.

V. Super-Stability and Anti-Scepticism

Is the Super-Stability Limitation implausible? If so, then the expressivist needs some way to make sense of the possibility that a belief is super-stable (inaccessible) and yet in error (true), an unenviable task within the present framework. But in fact I think the expressivist is better off digging her heels in.

The expressivist can make sense of the possibility that any of her beliefs (including regarding standards) are in error. This encourages a healthy kind of open-mindedness: it allows us to always be open to finding out that we’re in the wrong. There is, however, something distinctively *optimistic* about how the expressivist makes sense of this: it is in terms of *possible incremental self-correction*. (Note: not that her beliefs *must* or *will eventually* self-correct, but that they *can*.) That is: making sense of the very possibility of error uses the *anti-*

[21] When talking about *series* of improvements, we are here talking about what Ridge (2015) calls ‘EACH STAGE STABILITY’: *each* improvement in the series has to be endorsed by the (other) standards endorsed at *that* time.

sceptical idea that the truth is accessible in a certain way. As Blackburn explicitly argues in another context:

> ‘What is not guaranteed by this kind of thought [i.e., that any of our beliefs could be wrong] is the intelligibility of a different, radical kind of global error: the possibility that the truth might be nowhere that we can get to from here… [T]he idea that moral truth may be entirely and totally hidden from even our best efforts at improvement is not guaranteed to be coherent by reflection on those efforts and their structure.’ (1996: 94)

It is tempting to read this as an endorsement of the Super-Stability Limitation. When Blackburn mentions places “we can get to here”, he is of course not speaking merely causally – we might get anywhere from anywhere, causally speaking. He is talking about where I might get to via (self-endorsed) improvement from my present belief set. On its own terms, the Limitation amounts to a kind of anti-scepticism: any true moral belief is in principle accessible to me through self-endorsed improvement, and any false moral belief is in principle eliminable from my moral outlook through self-endorsed improvement. (Recall this is *actual* improvement by the lights of the standards I endorse and their improvements.) One might wish to reject such anti-scepticism, but it is difficult to see what could push the expressivist to do so. Any argument that tries to point to a realm of moral facts that may lie forever beyond our discovery will gain little traction. Indeed, in subscribing to this, the expressivist joins a long anti-realist tradition of thinking that the moral truths are *epistemically constrained* – that they are not evident-transcendent or do not, in some sense, outrun rational acceptability.[22] This anti-scepticism is plausible enough, and highly plausible within the expressivist's anti-realist outlook – to the point where it is difficult to see how we could have a reason to reject such anti-scepticism unless we already had reason to reject the expressivist outlook *tout court*.[23]

The re-vamped version of Egan's worry, however, will be that a belief can be super-stable by my lights, and a contrary belief super-stable by someone else's – meaning I'll be arbitrarily (and "smugly") committed to the epistemic priority of my own position. We will return to this worry below. But we will be able to get a much firmer handle on this issue, and thus how the expressivist can respond, only after we've chased down more of the implications of the expressivist theory of fallibility.

VI. The Weak Super-Stability Limitation

To proceed, I'm first going to articulate and motivate a further limitation on first-personal fallibility by expressivist lights: the Weak Super-Stability Limitation (VI). I'll then consider and reject the natural way to try and evade this limitation (VII-VIII).

[22] For an excellent overview of epistemically constrained theories of truth, see Künne (2003: ch.7). There is a difference between thinking that the moral *truths* are epistemically constrained and thinking that moral *truth* is epistemically constrained. Still, the latter can explain the former, so this gives the expressivist a further reason to endorse the theory of truth articulated below.

[23] By Lenman's (2014: 240) lights, Egan's "smugness" is a vice of excess to which there is an antipodal vice of deficiency: 'a kind of moral pusillanimity, a catastrophic lack of confidence in one's own moral convictions and commitments.' I like to think of this anti-scepticism as the virtue between Lenman's vices.

Suppose first that our agent has moral belief b_v: $b_v \in B_a$. But there is some improving change she may go through such that she no longer has that belief: there is some $B > B_a$ such that $b_v \notin B$. For instance, if she were more informed about the environmental impact of animal farming, she'd change her mind about vegetarianism. Now, this is *precisely* the situation that she is worried about when she worries about her fallibility. Is it, then, *a priori* (for her) that if a belief of her own is *unstable* in this way, then it is in error? This would be a major cost, for it would commit the expressivist to every first-personal self-endorsed improvement being an improvement *simpliciter*. Furthermore, if she were to go through the relevant improving change, she would *then* be able to make sense of the possibility that she is ignorant. It would thus be highly uncomfortable, perhaps incoherent, to maintain that she cannot *at present* make sense of this possibility. (This kind of *diachronic inconsistency* will be prominent in what follows.) So, the expressivist needs some way of making sense of an unstable belief being correct.

The natural recourse is to *further* self-endorsed improvement. Even if there is some $B > B_a$ such that $b_v \notin B$, there may yet be some $B' > B$ such that $b_v \in B'$. Diachronic consistency supports this suggestion: if she were to go through the relevant improving change from B_a to B, this is precisely how she would then make sense of the possibility of her ignorance.

However, the dialectic then repeats: what if there *is* such a belief set, B'? Is our agent then *at present* unable to make sense of the possibility that b is in error? Of course not, for there may be some $B'' > B'$ such that $b \notin B''$. And even if there is such a B'', there may be a $B''' > B''$ such that $b \in B'''$. And so on.

But this has a limit. For what if there is some $B > B_a$ such that b is super-stable at B? The Super-Stability Limitation says that, if she were to go through the relevant improvements, *then* the agent wouldn't be able to make sense of the possibility that b_v is in error. The new question is if she can make sense of this *at present*. And it is clear that she cannot use *further* self-endorsed improvement to do so. Likewise if there is some $B > B_a$ such that b_v is inaccessible at B. She would then be unable to make sense of the possibility that b_v is correct; and she cannot use further self-endorsed improvement to make sense of it at present either.

Label a belief that is super-stable at some $B' \geq B$, 'weakly super-stable' at B itself; while if a belief is inaccessible at some $B' \geq B$, then it is 'weakly inaccessible' at B. Further, let's re-label super-stability and inaccessibility, 'strong super-stability' and 'strong inaccessibility' respectively. The question is if an agent can make sense of the possibility that a belief that is weakly, but not strongly, super-stable at her present belief set is nonetheless in error; and likewise for a belief that is weakly, but not strongly, inaccessible at her present belief set being correct. As the preceding makes clear, she cannot do so through *further* self-endorsed improvement after the point where it is (strongly) super-stable or inaccessible. Label the contention that the expressivist cannot make sense of these possibilities, 'the Weak Super-Stability Limitation'.

VII. Divergent Improvements and Fallibility

If she needs to reject the Weak Super-Stability Limitation, our agent needs some way of making sense of the fallibility of those $B > B_a$ other than further self-endorsed improvement. The natural suggestion here is to appeal to *divergent* opinions among those $B > B_a$. There will, after all, be more than one way for our agent to improve her beliefs. Perhaps if our vegetarian became better informed about the environmental impact of animal farming she'd change her mind, but if she became better informed about the living conditions of livestock, she wouldn't. In general, it may be that while $b_v \in B1$ for some $B1 > B_a$, $b_v \notin B2$ for some B2 > B_a. And this is so even if b is strongly super-stable at B1 or strongly inaccessible at B2. So, perhaps our agent can at present make sense of the possibility that b_v is in error even if b_v is strongly super-stable at some B1 > B_a in terms of the possibility that there is some B2 > B_a such that b ∉ B2. (Let's focus on weak super-stability for the time being; the considerations run *mutatis mutandis* for weak inaccessibility.)

But things cannot be as simple as this. First, even if there is some such B2, B1 may be accessible via self-endorsed improvement from B2: even if B1 > B_a and B2 > B_a, it may be that B1 > B2. And since we make sense of the fallibility of B2 in terms of those B > B2, and B1 is one such, B2 then cannot be a sensible way to make sense of the fallibility of B1.

In general, for any two belief sets B1 and B2, there may be some B3 such that B3 ≥ B1 and B3 ≥ B2. Label any such B3 a 'point of convergence' for B1 and B2. At the limit, B3 may be B1 or B2 itself. The point of the last paragraph is that, if B1 is a point of convergence for B1 and B2, then B2 is a poor way to make sense of the fallibility of B1. If, on the other hand, B2 were a point of convergence for the two, then it would be a sensible way to make sense of the fallibility of B1; but this is just the "further self-endorsed improvement" understanding of fallibility that we're trying to find an alternative to. If b_v were strongly super-stable at B1, for instance, then if B2 is a point of convergence for B1 and B2, b_v is strongly super-stable at B2 too.

Suppose instead, then, that there is some point of convergence for B1 and B2, B3, that is distinct from both. Either b_v ∈ B3 or b_v ∉ B3. If b_v ∉ B3, then it disagrees with B1, and thus offers a way to understand its fallibility; but, again, this is just the "further self-endorsed improvement" understanding of fallibility we're trying to find an alternative to. If b_v is strongly super-stable at B1, for instance, then b_v is strongly super-stable at B3 too. If b_v ∈ B3, then it agrees with B1, and hence offers no way of understanding its fallibility. (There may, of course, be some series of self-endorsed improvement from B2 to some B4, distinct from B3, such that $b_v \notin$ B4. But then the dialectic reiterates with regards to B1 and B4, instead of B1 and B2.) So, if there is any point of convergence between B1 and B2, B2 offers no new way of understanding the fallibility of B1.

Therefore, this proposal requires that there is no point of convergence between B1 and B2. For any B1 and B2, if they have no mutually assessible point of convergence, then we'll label B1 and B2 'truly divergent'; if there is some point of convergence, then B1 and B2 are 'merely divergent'. If our agent is to find a new way to make sense of the fallibility of some B1 > B_a in terms of some B2 > B_a, then B1 and B2 need to be truly divergent.

But even this is not enough: $b_v \in$ B1 and $b_v \notin$ B2, but even so there may be some B3 > B2 such that $b_v \in$ B3. Again, since this is how we make sense of the fallibility of B2, the existence of B2 would be a silly way to make sense of the fallibility of B1. (There may, again, be some divergent B4 > B2 such that $b_v \notin$ B4, but then the dialectic reiterates with regards to B1 and B4, rather than B1 and B2.) So, what we require is that there is no B3 > B2 such that $b_v \in$ B3. (Or, more precisely, even if there is some such B3, there is some B4 > B3, such that $b_v \notin$ B4 and there is no B5 > B4 such that $b_v \in$ B5.) That is, we require that b_v is *strongly inaccessible* at B2.

The conclusion: divergent improvements from B_a only offer our agent a new way of understanding the fallibility of those B > B_a if they are *truly divergent* from one another, and are maximally stable in their disagreement: if b_v is strongly super-stable at one, and strongly inaccessible at the other.

VIII. Problems with True Divergence

This constitutes a choice-point. Either we admit that this scenario is possible – that there may be some B1 > B_a such that b_v is strongly super-stable at B1 and some B2 > B_a such that b_v is strongly inaccessible at B2 – or we give up on avoiding the Weak Super-Stability Limitation in this way. The latter strikes me as the preferable option. I will not make a decisive case, but there are clear motivations for rejecting the former. Here are five.

First, suppose the possibility is actualised. How can the expressivist then make sense of the idea that one of B1 or B2 is in the right, the other in the wrong? After all, the account of fallibility applies symmetrically: she at present makes sense of the possibility that B1 is in error in terms of the possible existence of some B2, and likewise the fallibility of B2 in terms of some possible B1. So if this possibility is actualised, what then can she say? That neither is in the right? That would require finding space between being true and not being true. That both are in the right? Even granting dialethism, this means that there is no fallibility after all.[24] These options aside, we need a tie-breaker, and we do not at present have one. How, then, can the expressivist even *make sense* of the idea that one is better off than the other?

Furthermore, suppose for illustrative purposes that B1 is in the right and B2 is in the wrong. B2 is then in a sceptical scenario: there is no series of self-endorsed improvement one can go through from B2 such that one comes to believe the truth with regards to b_v. In conjunction with the Super-Stability Limitation, this introduces at least two complications. On the one hand, it introduces a kind of diachronic inconsistency into the account. Our agent can *at present* make sense of the possibility that B2 is in the wrong, but would not be even be able to make sense of this possibility if she went through the relevant improvements and wound up at B2.

[24] For a gesture at a potential way to resolve this via endorsing a kind of relativism, see fn.36.

On the other hand, and relatedly, the Limitation articulates our agent's commitment to a certain kind of optimism when it comes to her epistemic standards: they are capable of leading her out of error. But allowing truly divergent improvements from B_a means allowing that they can *also lead her into* exactly the sceptical scenario she cannot make sense of being in at present. There is a clear tension here: it is not obvious that one can coherently endorse one's standards *as* standards while allowing that they can lead one so astray.

Moreover, B2 is, by her own lights, an *improvement*: if she occupies B2 she is better off, epistemically speaking, than she is at present. While we should allow that B2 may be mistaken about things that B_a is not, it is far from obviously coherent to allow that someone that is by your own lights in a better epistemic situation to you somehow manifests a sceptical possibility you cannot make sense of being in yourself.

Finally, building on this last point, this modification to the account of fallibility actually introduces an uncomfortable *bifurcation*: *for those* $B > B_a$, fallibility is not only understood in terms of self-endorsed improvement, but using belief sets accessible via divergent series of self-endorsed improvement from B_a. But if there is such a bifurcation for those $B > B_a$, should there not likewise be one for B_a itself? Otherwise we risk arbitrarily privileging our own starting points – indeed, doing so over belief sets that are actually *improvements* on our own by our very own lights, bringing about the tension just discussed. But it is unclear what the second conjunct should be: it cannot be the existence of just any belief set that disagrees with B_a, that was the whole point of making use of the idea of *improvements* in the first place. Perhaps we could appeal to belief sets truly divergent[25] from B_a that are accessible via self-

[25] We need true divergence here for the same reasons we needed in the last section.

endorsed improvement from belief sets from which B_a is also so accessible. But I suspect this notion is ultimately too unconstrained to be of use.

I don't claim that these worries are decisive, but I do suspect that placating them would require formulating a whole new way of making sense of fallibility within the expressivist framework. At least given the Blackburnian approach we're interested in here, then, rejecting the possibility of such truly divergent belief sets among those $B > B_a$ strikes me as the preferable option. The above discussion should also make clear that it's not obvious how *else* one might avoid the Weak Super-Stability Limtiation without formulating a different account of fallibility.

But another lesson is that taking on this limitation is no great cost, or at least no great *additional* cost given the prior commitment to the (Strong) Super-Stability Limitation. Indeed, the preceding five worries suggest that it would be quite uncomfortable – potentially even incoherent – for our agent to allow that those $B > B_a$ are vulnerable to a kind of error to which B_a itself is not vulnerable.

The main cost, it seems to me, would arise if we could be persuaded that true divergence is a reasonably widespread phenomenon: that it is not unusual to find two belief sets that share no potential point of convergence through self-endorsed improvement. For then rejecting this possibility within those $B > B_a$ might be seen as mere wishful thinking. But in fact I think the opposite is true. It is to this topic, and the updated version of Egan's worry, that we now turn.[26]

[26] Horgan & Timmons (2015: 197) develop a proposal kindred to my own, and seem to be committed (2015: 202) to the analogue of the Weak Super-Stability Limitation in their framework. However, they do not motivate or address the plausibility of this limitation, nor the possibility of divergent opinions being weakly

IX. Weak Super-Stability and Moral Networks

I have argued that the expressivist is committed to weak super-stability (by her own lights) being sufficient for truth; and, by parity, weak inaccessibility being sufficient for untruth.[27] Can a belief, b, be neither weakly super-stable nor weakly inaccessible for an agent? This requires that, for every $B \geq B_a$ such that $b \in B$, there exists some $B' > B$ such that $b \notin B_a$; and for every $B \geq B_a$ such that $b \notin B$, there exists some $B' > B$ such that $b \in B'$. That is, no matter how much she improves her beliefs, there is some improving change she can go through such that she changes her mind about b, and a further such change through which she changes her mind again, and so on. Call such a belief, a *flip-flop* belief. Where super-stable and inaccessible beliefs have an enduring stability, flip-flop beliefs have an enduring instability. As such, they alone present many of the same problems that true divergence among those $B > B_a$ would present. For instance, it seems b must be either true or untrue, but it is difficult to see how the expressivist can make sense of either possibility, given her

super-stable (in their terms, "I-stable" along different "I-trajectories"), which have been my preoccupations in sections IV-VIII. I also find their response to the smugness worry unsatisfying – see fn.38. However, while I have certain misgivings about the constraints they put on "I-trajectories" (the analogue of series of self-endorsed improvements), if one prefers their underlying framework, I see no reason to doubt that the considerations of sections IV-X carry over to it. I thus see our proposals as close allies.

[27] Why untruth rather than falsity? It seems that judging that eating meat is not wrong is distinct from simply not judging that eating meat is wrong. So, *both* the judgement that eating meat is wrong *and* the judgement that eating meat is not wrong may be strongly (and hence weakly) inaccessible at some belief set. Thinking of falsity as truth of negation, then, weak inaccessibility is insufficient for falsity.

account of fallibility: the situation she is worried about *qua* fallibility is *always* realised. Since this runs for both being mistaken *and* being ignorant, absent some tie-breaker it's not clear how we can make sense of *either* possibility. Perhaps, then, the expressivist has equal reason to rule out the possibility of flip-flop beliefs among those $B > B_a$.[28]

Setting these aside, any belief that is not weakly super-stable at a belief set is weakly inaccessible, and *vice versa*. Weak super-stability is thus *necessary and sufficient* for truth, and weak inaccessibility *necessary and sufficient* for untruth. We have arrived at necessary and sufficient conditions on moral truth in the form of weak super-stability.

Our updated version of Egan's "smugness" worry will arise from the fact that a belief might be weakly super-stable for one agent, while a rival belief is weakly super-stable for another: e.g., that b is strongly super-stable at some $B_a+ \geq B_a$, while ¬b is strongly super-stable at some $B_b+ \geq B_b$. Assuming that where there is moral disagreement there is error, at least one of the beliefs must be in the wrong. Note that this does not entail that one of the *agents* is *at present* in the wrong: since *weak* super-stability is about which (strongly super-stable) beliefs are access*ible* to an agent, it may be that e.g. $b \in B_a$ and $b \in B_b$. Rather, each agent is committed to thinking that the other can *arrive* at a strongly super-stable, but mistaken, belief through self-endorsed improvement, while they themselves cannot; each allows that the other's standards can lead them into a sceptical scenario, but cannot make sense of the suggestion that their own standards can do so. And this looks arbitrary.

[28] Blackburn (1993: 22) considers such a case and comes very close to rejecting the possibility of flip-flop beliefs. I think this is probably the right option. However, I'm intrigued in the possibility that flip-flop beliefs offer an expressivist spin on the idea that a concept is "inconsistent". The liar sentence, for instance, might be thought to express a flip-flop belief; the instability is highly reminiscent of Gupta & Belnap's (1993) "revision" theory of truth, for instance. But this is too large a topic to broach here.

Now, there is a potential asymmetry here. But the *force* of this worry depends on two things: (a) there being beliefs sets that differ with regards to which beliefs are weakly super-stable at them, or are *WSS-divergent*; and (b) *which* belief sets are WSS-divergent. With regards to (a), B_a and B_b are WSS-divergent iff there is some b such that: (i) b is weakly super-stable at B_a but not at B_b, or (ii) b is weakly super-stable at B_b, but not B_a. If there is no WSS-divergence, there is no problematic asymmetry. We can get a better handle on WSS-divergence, and hence the worry, by staying at a level of formal abstraction for the time being. We turn to (b) in the next section.

WSS-divergence, it transpires, is intimately tied up with true divergence. (The following two paragraphs can be skimmed by those only interested in the upshot of these results, which is discussed in the following paragraphs.)

First, if b is weakly super-stable at B_a but not B_b, then B_a and B_b are truly divergent. For suppose, for *reductio*, that they are merely divergent. Then they have some point of convergence, B1. Since b is weakly super-stable at B_a, there is some B_a+ such that B_a+ ≥ B_a and b is strongly super-stable at B_a+. Since both B_a+ ≥ B_a and B1 ≥ B_a, and there is no true divergence among those B > B_a, B_a+ and B1 must have some point of convergence, B2; and since b is strongly super-stable at B_a+ and B2 ≥ B_a+, b is strongly super-stable at B2. As B2 ≥ B_b, b is then weakly super-stable at B_b, which we are assuming it is not.

We could run the same proof the other way if our agent could assume that there is no true divergence among those B > B_b. But while I have argued that this is a plausible commitment for an agent to have regarding her own belief set, I have not argued that it is a plausible commitment to have regarding *every possible* belief set. Some may endorse some quite pathological standards after all. However, we can prove that, if b is weakly super-stable at B_b but not B_a, then *there is some B ≥ B_b* that is truly divergent from B_a. For if b is weakly super-stable at B_b, there is some $B_b+ \geq B_b$ where b is strongly super-stable. Suppose, then, for *reductio* that B_b+ and B_a are merely divergent, so they have a point of convergence, B1. Since $B1 \geq B_b+$, b is strongly super-stable at B1; and since $B1 \geq B_a$, b is therefore weakly super-stable at B_a, which we're assuming it's not. So, B_a and B_b+ are truly divergent.

This means that our agent's belief set falls within a *network* of moral outlooks where all and only the same moral beliefs are weakly super-stable. In general, two belief sets B1 and B2 are part of the same network iff: (i) B1 and B2 are merely divergent, and (ii) for any B1+ > B1 and B2+ > B2, B1+ and B2+ are merely divergent. Since there is no WSS-divergence within a network, there can be no problematic asymmetry arising from the weak super-stability limitation between belief sets in the same network.

Call our agent's network, N_a. Since there is no problematic asymmetry within N_a, our agent can only be accused of considering herself privileged over those outside her network; i.e., those who fail condition (i) or (ii) for B_a.

As far as the charge of smugness goes, I do not think that those belief sets that only fail condition (ii) will pose any problem not posed in more profound form by those that fail condition (i). Those that fail condition (i) are truly divergent from B_a's network. By B_a's lights, then, there is some truth that *cannot* become strongly super-stable for them through self-endorsed improvement, while that is not so for B_a. Those that *only* fail condition (ii), however, are merely divergent from B_a; since N_a is thus accessible via self-endorsed improvement, by B_a's lights there is no truth that is *actually* inaccessible to them in this way. Rather, there is *also* some belief set (and hence network) accessible to them that is truly divergent from all those belief sets in N_a, including B_a, i.e., fails condition (i). So, the asymmetry is that it is *possible* that a truth should become inaccessible to them, in the way described, through self-endorsed improvement, while this is not so for B_a. This is a less radical asymmetry. For ease of presentation, then, we shall focus just on the potential asymmetry with those that fail condition (i), i.e., that are truly divergent from B_a. If we can placate the charge of "smugness" here, this should carry over to those that only fail condition (ii).

X. Smugness Towards Other Moral Networks

The issue, then, turns on which belief sets are WSS-divergent from B_a, and since condition (i) of WSS-divergence requires true divergence, our question becomes: which belief sets are plausibly truly divergent from B_a? And in answering this, we should immediately realise that not just *any* belief set being truly divergent from B_a is sufficient to establish a *problem*. Our agent, we can assume, endorses sensible standards like Blackburn's information, sensitivity, maturity, imagination, and coherence. Now imagine, if you can, someone who endorses no standards whatsoever. This person exists in a network of one. Or consider someone who endorses silly standards like *misinformation*, *insensitivity*, *immaturity*, *close-mindedness*, and *incoherence*. If there can be such agents, then they may well be truly divergent from B_a. But it is difficult to feel at all *troubled* by this. "Smugness" here amounts to saying that there is some moral error that it is impossible for them to ever *permanently* expunge through self-endorsed improvement, or some true moral belief that is impossible to *permanently* instil through self-endorsed improvement. (Note this is permanence *through self-endorsed improvement*. It's still possible for any such agent to, e.g., become part of our agent's network through changes he doesn't so endorse.) This is, if anything, a *plausible* prediction of our theory.[29]

To generate a problem, then, our objector needs to give us reason to think that there is some – for want of a better word – *reasonable* moral outlook that falls outside of our agent's network. If two perfectly reasonable, but nonetheless quite fundamentally opposed agents – like, perhaps, the consequentialist and the deontologist, or maybe the conservative and the liberal – existed in different networks, this would start to look like a problem. Call this, a *Reasonable, In-Principle Irresolvable Dispute*, or RIPID.[30] RIPIDs are engendered by true divergence between reasonable agents. However, in what follows, I argue first that admitting the possibility of RIPIDs, and thus the charge of smugness, requires pushing our intuitions concerning the potential robustness of disagreements between reasonable moral agents to perhaps implausible extremes. I then point out that, even if we're willing to allow that there may be RIPIDs, we also require absolutism to create a problem for expressivism, and that this combination is also problematic for representationalists.

[29] Compare e.g. Blackburn (1984: 199) on 'the vague and unfounded disquiet that I have no right to judge unfavourable people with any other opinion'. It is not the *mere* existence of other outlooks that ought to be troubling, but outlooks of a certain *quality*. The reader will note strong resonance between the issues here and Blackburn's (1984: 197-202) discussion of a 'tree… [where] each node (point at which there is branching) marks a place where equally admirable but diverging opinion is possible' and how we might '*transcend the tree structure*' – though there are key differences too. Indeed, at every stage in this paper I have been heavily influenced by Blackburn's work, and the proposal is in this sense thoroughly Blackburnian. Nonetheless, I doubt very much that (at least present-day) Blackburn would have any time for the resultant view.

[30] Note that such disagreements are not "in principle" irresolvable in the sense that the agents cannot come to agree, but that they cannot do so *through self-endorsed improvement*.

There is perhaps a widespread intuition that there can be very trenchant disagreements between reasonable moral agents; and indeed, it may seem that the expressivist is well-placed to explain and embrace this possibility.[31] Initially, then, denying that there are any RIPIDs may seem surprising. However, we can accommodate this intuition without *also* conceding that there are RIPIDs; that the ordinary consequentialist and deontologist, for instance, might be truly divergent from each other.

In the first place, note that there is an inverse relationship between the plausibility that a belief set is truly divergent from B_a and how problematic our agent's smugness is. True divergence becomes more plausible with more radical difference in epistemic standards. But as our agent's interlocutor's standards differ from sensible things like information and coherence, we become closer to the silly examples mentioned above, and our agent's smugness becomes quite understandable and tolerable.

Suppose, then, that both our agents endorse sensible standards like information, coherence, and so on. For the case to constitute a RIPID thus requires that there is *no possible agent* whose moral outlook is more informed, more coherent, etc. than both the consequentialist's and the deontologist's. Crucially, note that saying that there *is* some such agent does not make the interlocutors' agreement through self-endorsed improvement *inevitable*, for *any* such improvement may be (merely) divergent from this point of convergence. It only makes potential convergence *possible*. (Also note that thinking that there could be such an agent does not commit one to thinking that *this* point of convergence gets all the ethical questions right. This improved agent may further improve their beliefs;

[31] Thanks to an anonymous referee for pressing me on this.

and even if there is one such agent that is a consequentialist, there could be another that is a deontologist.)

We can therefore allow that the ordinary consequentialist and deontologist are in fact incapable of *convincing* one another, no matter how ingenious their arguments. After all, their opponent may respond with a yet more sophisticated argument; and, in any case, humans are stubborn and limited, and subject to all kinds of cognitive biasing effects. So, their disagreement might be *in practice* irresolvable. More importantly, we can allow that they can each become more informed, sensitive, coherent, and so on while each holding on to their fundamental ethical views. Indeed, we can even allow that it's in principle possible that they should each *continuously* improve their views – endlessly becoming more informed, more sensitive, more coherent, etc. – without *ever* coming to an agreement.

This, I submit, straightforwardly accommodates the force of the intuition that there can be highly robust disagreements between reasonable moral agents. A disagreement does not have to be a RIPID for the agents to have no guarantee that they will come to an agreement, no matter how long and hard they may work at it.

Allowing the possibility of RIPIDs between ordinary moral agents requires taking on an extra commitment, pushing our intuitions one step further, to extremes that one might consider quite implausible.[32] For it requires ruling out that there is *any possible agent* that is

[32] Note that when Egan talks about smugness, he talks about "fundamental moral disagreement". I have avoided talking in any such terms. There is an everyday sense of this phrase in which fundamental moral disagreements are undoubtedly common enough occurrences between ordinary agents, perhaps even quite reasonable ones. I think the phenomenon Egan is really interested in is rarer. As I hope this discussion makes clear, RIPIDs certainly are.

more informed, more sensitive, and more coherent than both of the agents involved. Allowing RIPIDs *between ordinary, reasonable moral agents* thus looks like it requires denying not just that there is, but that there *could be* a God, an ideal moral agent,[33] or even just a kind of being substantially epistemically better-off than humans. An alien species capable of moral judgement but with substantially improved information-processing power, emotional sensitivity, intellectual honesty, reasoning capabilities – if this is merely possible, then such a specimen could effectively deploy its resources to formulate a moral outlook that is more informed, sensitive, coherent, etc. than any human's.[34] It would hereby constitute a point of convergence for our interlocutors. (And, to reiterate, that's *not* for one second to say, "and is therefore right"!) It is a substantial and contentious thesis that such a being is impossible!

One may think that it is one thing to rule out RIPIDs between ordinary moral agents, quite another to rule out RIPIDs *tout court*. Couldn't there, e.g., be a RIPID between two more advanced moral agents, perhaps two beings who are substantially epistemically better-off than any human, one of whom is a consequentialist, the other a deontologist? But given that our agent rules out any true divergence among those $B > B_a$, the problem is in seeing how any two belief sets could *both* be more advanced moral agents – i.e., improvements on B_a – *and* truly divergent from each other.

[33] The possible existence of a God or ideal moral agent is hereby sufficient, though not necessary, to eliminate the smugness problem. While such an assumption is contentious, it is one often found in other metaethics and theories of truth. The denial of RIPIDs is a weaker thesis, and thus necessarily more plausible.

[34] For ease, I talk here as though standards like these are the correct ones to endorse, but a similar point also runs for any set of improved standards. This only makes the point stronger, since it broadens the potential points of convergence.

Suppose our ordinary human interlocutors have belief sets B_a and B_b. They have a point of convergence, B_c; and suppose B_c is truly divergent from some B_d. The question is whether or not the disagreement between B_c and B_d could be a RIPID. I've argued that B_a ought to rule out that there is any true divergence among those $B > B_a$. Therefore, since $B_c > B_a$ and B_c and B_d are truly divergent, it's not the case that $B_d > B_a$. Moreover, B_d and B_a must be truly divergent. For suppose for *reductio* that B_d is merely divergent from B_a, so there is some point of convergence B_e, such that $B_e > B_a$ and $B_e > B_d$. As $B_e > B_a$ and $B_c > B_a$, B_e and B_c must be merely divergent. But if B_e and B_c are merely divergent and $B_e > B_d$, then B_c and B_d are merely divergent, which we're supposing they're not. So, B_a and B_d are truly divergent. Therefore, not only is B_d *not* more informed, sensitive, coherent, etc. than B_a, neither is *any* belief set accessible from B_d via self-endorsed improvement. But then we've lost all sense in which B_d could be a more advanced moral agent than B_a; and it is difficult to see how B_d could endorse sensible standards like information, sensitivity, coherence, etc.[35] As emphasised at the beginning of this section, the mere possibility of some belief set truly divergent from B_a is insufficient to establish a problem for the expressivist.

While I do not claim that these considerations are decisive, maintaining that RIPIDs are possible is hereby shown to be a highly contentious commitment, which the expressivist can reasonably reject. If there are no RIPIDs, then there is no problem: even though there can be disagreements between reasonable agents potentially irresolvable through endless series of self-endorsed improvements, all reasonable moral agents exist within a vast network

[35] One might suggest that B_d is truly divergent from B_a because they are *similarly* informed, sensitive, coherent, etc., and there is no point of convergence. But then this is just another case of a putative RIPID between ordinary moral agents. (Also, note that, if $B_d > B_b$, then this is a case of B_b falling outside B_a's network in virtue of failing condition (ii).)

in which all and only the same beliefs are weakly super-stable. Rejecting RIPIDs provides a straightforward and conclusive response to the charge of smugness arising from potential WSS-divergence, and is therefore the response I find most attractive.

For the sake of completeness, however, suppose that we're convinced that there can be RIPIDs: reasonable moral agents like the consequentialist and deontologist can exist in distinct moral networks such that there is no belief set more informed, sensitive, coherent, etc. than both of them – their disagreement is not just possibly irresolvable, but necessarily so. Whether or not we can hereby pin a charge of smugness to the expressivist still turns on the further question of whether we are relativists or absolutists.

That there may exist irresolvable moral disputes of one kind or another is a familiar contention of metaethical relativists. It is, of course, a matter of contention to what extent enduring moral disagreement supports relativism, but let us grant the relativist thesis for the time being. Now, short of there being any special *problem* for the expressivist, the idea of weak super-stability within a network actually provides a worked-out and sensible way of understanding the relativist thesis: granting both relativism and the possibility of RIPIDs, the sensible thing to say is that moral truth is relative, in one way or another, to your moral network.[36]

[36] Allowing the relativist thesis here opens up some new choice-points for the expressivist. For instance, suppose that there is some $B1 > B_a$ and $B2 > B_a$ such that B1 and B2 are truly divergent, so two different moral networks are accessible to our agent. We might then say that the moral truth is indeterminate for this agent for any moral *p* on which the two networks disagree. As my own inclination is absolutist, and in any case the relativist options here strike me as less plausible then those discussed in the main text, I resist discussing these options in any more detail for the sake of space.

No doubt there are serious questions about how to understand this relativism. For one thing, *qua* moral agent, each of us looks committed to our own network being correct, and hence the incorrectness of those networks that disagree with it. *Being a relativist* seems to require occupying a neutral, "God's-eye" perspective on the debate that we do not in fact have. Similarly, we presumably want there to be constraints on *which* networks get their own moral truths – for, as discussed, some are bound to be quite pathological. But the idea that the different networks all meet some standardised set of standards is in tension with the idea that there is true divergence between them after all. But these are complications introduced by the *relativism*, not the *expressivism*.

The charge of smugness can only get any grip at all, then, given the combination of RIPIDs and absolutism. In that case, at most one of the moral networks gets the moral facts right.[37] *Qua* moral agent herself, each expressivist is committed to her own moral network being the lucky one. And this looks arbitrary, hence the charge of smugness.

But the combination of RIPIDs with absolutism is an unhappy one. For suppose that one is a representationalist instead. One is still committed to thinking that at most one of the moral networks gets the moral facts right. And, *qua* moral agent oneself, you are committed

[37] An interesting and tempting way of trying to combine RIPIDs with absolutism would be to hold that 'p' is determinately true iff the judgement it expresses is weakly super-stable in *every* (reasonable) moral network; 'p' is determinately untrue iff it is not weakly super-stable in any network; and otherwise it is indeterminate whether 'p' is true or untrue. The thought being that one is perhaps not *mistaken* to think that p (or not think that p) if it is indeterminate whether 'p' is true or untrue. This would be unpalatable if it rendered moral indeterminacy rampant, but recall that we're considering *reasonable* true divergence here: the expressivist might argue that RIPIDs are thus sufficiently rare to only introduce indeterminacy in a tolerable number of recherché cases. But moral indeterminacy raises a number of questions I cannot get into here.

to your own network being the lucky one. Both the expressivist and the representationalist, then, think that they are in an epistemically privileged situation: there is some truth that will not become weakly super-stable for someone in a rival moral network, no matter how much more informed, sensitive, coherent, etc. they become; but this is not so for those in their own network. The difference between the two is that the representationalist can *make sense* of the possibility that it is his own moral network that is mistaken. For his explanatory story *starts* with moral facts, which our moral judgements aim to represent. Perhaps, then, the beliefs that are weakly super-stable within his own network get the facts wrong. The representationalist can "step outside" his first-order ethical views, when in the metaethics classroom, to make sense of this possibility. The expressivist, it seems, cannot.[38]

However, the representationalist faces his own problems. For, at first pass, making sense of the possibility that his own network is in the wrong comes at the expense of ever *knowing* which network is in the right. While the representationalist doesn't directly use standards to explain fallibility, he presumably still thinks that our *access to* the moral truths is in some sense mediated by the epistemic standards he endorses. Granting that there are moral facts that our moral judgements aim to represent, how else could we hope to find out what these moral facts are than by becoming more informed about non-moral matters, more

[38] In considering the charge of smugness, Horgan & Timmons (2015: 202-3) point to the recent debate concerning the rational response to disagreement with epistemic peers, where many hold that the thing to do is to retain one's own opinion. If this asymmetric privileging of one's own opinion is rational, it can hardly be objectionably smug; and, as they emphasise, this is a general view, not peculiar to expressivism. However, I think this misses the thrust of the objection. Even if this is the *rational* thing to do, one can still *make sense* of the possibility that you are the one in the wrong; and the expressivist rules this out – or so the objection goes. It is this stronger challenge that I'm responding to in sections IX-X.

sensitive, more mature, more imaginative, and more coherent? The moral truth, then, lies forever beyond the grasp of at least one of the networks (in the specified sense), and *which* one of them it lies beyond the grasp of lies beyond the grasp of both of them.

So while the representationalist can make sense of their own potential fallibility in a way the expressivist cannot, this seems to come with a sceptical cost. (And these twin worries for absolutism might be thought to motivate an 'if RIPIDs, then relativism' conditional.) Now, one might flat-footedly press that we *can* make sense of this possibility, and thus that expressivism is to be rejected. But here the expressivist can dig her heels in again. While I can say the words 'the rival network may be in the right', the expressivist says that the state this expresses is ultimately (though of course not obviously) incoherent.[39] (We might compare the person who thinks they can make sense of a triangle whose internal angles add up to 181 degrees, perhaps by imagining a diagram with the angles labelled '31', '58', '92'.) Barring this, there is remarkable parity between the expressivist and the representationalist.

To break this parity, the representationalist may try to appeal to some epistemic standard to which the expressivist is not entitled – some faculty of moral intuition or innate sensitivity to the moral facts. But this, the expressivist famously complains, is to solve a problem with a mystery. The expressivist's naturalistic aversion to 'divine sparks, skyhooks, faculties of intuition, cognitive powers beyond anything given by the five senses and general intelligence' (Blackburn 2010: 301) means they are unlikely to be moved by any such claim to dialectical advantage. These epistemological worries strike to the heart of representationalism.

[39] See Horgan & Timmons (2015: 205-6) for a different strategy, where their analogue of this sentence is understood as a meaningful (but false) piece of *metaethical* discourse, not first-order ethical discourse.

To pin a charge of smugness on the expressivist, then, requires taking on commitments that are – both individually and in combination – contentious, and which give rise to serious epistemological concerns for the representationalist. If we give up the absolutism, the expressivist has a sensible way of cashing out the resultant relativism. But the most attractive option, I have suggested, is to deny the possibility of RIPIDs, and hereby reject the charge of smugness outright.

XI. Conclusion

I have argued that, for the expressivist, the correct moral judgements are those that are weakly super-stable within her moral network. This allows us to give a substantive account of the truth conditions of atomic moral sentences. 'Eating meat is wrong' is true just in case the state that it expresses – disapproval of eating meat, say – is weakly super-stable within my network.[40] If we understand expressivism as a metasemantic thesis, then these truth conditions can provide the basis for a compositional, truth-conditional semantics for moral (and, of course, mixed) discourse. This is a major step forward: it establishes parity between the metasemantic expressivist and her representationalist opponent, by showing that the expressivist too is entitled to truth-conditional semantics.

But this is only useful if the theory of truth is plausible, and I have primarily been concerned to defend the plausibility of thinking that the moral truths coincide with the beliefs

[40] The indexical 'my' might cause discomfort here, but it should be understood *de re* and not *de dicto*. If some (non-contingent) moral judgement is weakly super-stable in my network – the network in which I happen to occupy a node in the actual world – then it is so at all possible worlds, even if I somehow occupy nodes in different networks in different worlds.

that are weakly super-stable within my moral network (partly, of course, on the grounds that this is your moral network too). While the discussion has been quite involved in places, the general picture is, I think, quite intuitive. While we all have quite different moral views, we generally endorse epistemic standards that are *capable* of bringing us to agreement. Fallibility is understood in terms of instability under improvement, and hence truth in terms of stability under improvement. This is a clear inheritor of the epistemically constrained theories of truth of the past, including coherence theories and the Peircean conception of truth as what is believed at the idealised limit of inquiry, as well as Crispin Wright's generalisation of mathematical proof, "superassertibility".[41]

One might worry that in endorsing a substantive account of moral truth, my version of expressivism has become too "realist". This strikes me as for the most part a book-keeping matter. Labels like 'realism' and 'anti-realism' only matter insofar as they indicate the kinds of *motivations* a view might have; so the worry has substance insofar as it is the worry that the details of my theory may betray its underlying motivations. This is a matter for another day. But I am not worried because our explanatory starting point was austere: we started only with atomic moral sentences expressing desire-like states and Blackburn's theory of moral fallibility; the rest of the work *shows* that we are *entitled* to a substantive theory of moral truth on this basis alone. Given the potential theoretical pay-off of such entitlement, I am convinced that this can only be a good thing.

[41] Roughly, a sentence is superassertible just in case it is assertible in some state of information that the world could generate in a suitably receptive inquirer, and would remain assertible no matter how that state of information was improved upon – see e.g. Wright (1992).

References

Baker, Derek and Woods, Jack. 2015. How expressivists can and should explain inconsistency. *Ethics*. **125**(2), pp.391-424.

Blackburn, Simon. 1984. *Spreading the word: groundings in the philosophy of language.* Oxford: Oxford University Press.

Blackburn, Simon. 1993. *Essays in quasi-realism*. Oxford: Oxford University Press.

Blackburn, Simon. 1996. Securing the nots: moral epistemology for the quasi-realist. In: Sinnott-Armstrong, Walter and Timmons, Mark eds. *Moral knowledge? New readings in moral epistemology*. Oxford: Oxford University Press, pp.82-100.

Blackburn, Simon. 1998. *Ruling passions: a theory of practical reasoning*. Oxford: Oxford University Press.

Blackburn, Simon. 2009. Truth and *a priori* possibility: Egan's charge against quasi-realism. *Australasian Journal of Philosophy*. **87**(2), pp.201-213.

Blackburn, Simon. 2010. Truth, beauty and goodness. In: Schafer-Landau, Russ ed. *Oxford Studies in Metaethics*, **5**(1), pp.295-314.

Chrisman, Matthew. 2012. On the meaning of 'ought'. In: Schafer-Landau, Russ ed. *Oxford Studies in Metaethics*, **7**(1), pp.304-332.

Dreier, James. 1996. Expressivist embeddings and minimalist truth. *Philosophical Studies*, **83**(1), pp.29-51.

Egan, Andy. 2007. Quasi-realism and fundamental moral error. *Australasian Journal of Philosophy*. **85**(2), pp.205-219.

Elstein, Daniel. 2013. *Prescriptions and universalizability: a defence of Harean ethical theory*. Ph.D. thesis, University of Cambridge.

Gupta, Anil and Belnap, Nuel. 1993. *The revision theory of truth*. Cambridge, MA: MIT Press.

Horgan, Terence and Timmons, Mark. 2006. Cognitivist expressivism. In: Horgan, Terence and Timmons, Mark eds. *Metaethics after Moore*. Oxford: Oxford University Press, pp.255-298.

Horgan, Terence and Timmons, Mark. 2015. Modest quasi-realism and the problem of deep moral error. In: Johnson, Robert and Smith, Michael eds. *Passions and projections: themes from the philosophy of Simon Blackburn*. Oxford: Oxford University Press, pp.190-209.

Köhler, Sebastian. 2014. Expressivism and mind-dependence. *Journal of Moral Philosophy*, **11**(6), pp.750-764.

Köhler, Sebastian. 2015. What is the problem with fundamental moral error? *Australasian Journal of Philosophy*. **93**(1), pp.161-165.

Künne, Wolfgang. 2003. *Conceptions of truth*. Oxford: Oxford University Press.

Lenman, James. 2014. Gibbardian humility: moral fallibility and moral smugness. *Journal of Value Inquiry*. **48**(2), pp.235-245.

Lewis. David. 1975. Languages and language. In: Gunderson, Keith ed. *Language, mind, and knowledge: Minnesota Studies in the Philosophy of Science*, **7**(1). Minneapolis, MN: University of Minnesota Press, pp.3-35.

Ridge, Michael. 2014. *Impassioned belief*. Oxford: Oxford University Press.

Ridge, Michael. 2015. I might be fundamentally mistaken. *Journal of Ethics and Social Philosophy*. **9**(3), pp.1-21.

Schroeder, Mark. 2008. *Being for: evaluating the semantic program of expressivism*. Oxford: Oxford University Press.

Schroeder, Mark. 2010. *Non-cognitivism in ethics*. London: Routledge.

Schroeder, Mark. 2014. Does expressivism have subjectivist consequences? *Philosophical Perspectives*. **28**(1), pp.278-290.

Smith, Michael. 2002. Evaluation, uncertainty and motivation. *Ethical Theory and Moral Practice*. **5**(3), pp.305-320.

Wright, Crispin. 1992. *Truth and objectivity*. Cambridge, MA: Harvard University Press.