---

**Article:**
Williams, JRG orcid.org/0000-0003-4831-2954 (2018) Normative Reference Magnets. Philosophical Review, 127 (1). pp. 41-71. ISSN 0031-8108

https://doi.org/10.1215/00318108-4230057

---

# Normative Reference Magnets

J Robert G Williams

**Abstract**
The concept of moral wrongness, many think, has a distinctive kind of referential stability, brought out by moral twin earth cases. This paper offers a new account of the source of this stability, deriving it from a metaphysics of content : « substantive » radical interpretation, and first-order normative assumptions. This story is distinguished from extant « reference magnetic » explanations of the phenomenon, and objections and replies considered.

## Part I: THE REFERENTIAL STABILITY PUZZLE.

Some say that the concept *moral wrongness* has a distinctive referential stability. Environment, society and moral opinion may vary dramatically, and yet agents will succeed in thinking about a common subject matter, right and wrong, so long as deliberation and sentiments are internally regulated in the right way. Schematically: there is a property P, applying to actions, which our concept of moral wrongness picks out, and there is also a conceptual role R which our concept of moral wrongness plays. The referential stability thesis is the following: necessarily if an agent has a concept W that plays role R, then W denotes P.[1]

Whether moral wrongness is really referentially stable in this way is a matter of dispute. This paper offers something to those on both sides of the issue. To those who endorse stability, I offer a theory of content that explains this puzzling phenomenon, and so by their lights a piece of evidence in favour of that theory of content. The explanation will be compatible with metaethical positions such as synthetic reductionism and non-naturalism, so demonstrates their compatibility with referential stability. Those who reject stability are free to read the derivation to follow as a reductio of the theory of content I develop. Either way, we make progress.

### 1.1 The stability thesis

[1] In fact, the statement will be refined in section 4.1.

Stability says: necessarily, if an agent has a concept W that plays role R, then W denotes property P. But what is the concept W? What role R features in the stability thesis? Which property P ends up denoted? I'll take these questions in reverse order.

The property P is the property of being morally wrong. In order to have a concrete thesis to work with, I take a particular candidate specification as a working assumption. I henceforth assume, with the Kantians, that P is the property *violating of the categorical imperative.*

R is the conceptual role of wrongness. A (conceptual) role is a possible pattern in the causes and causal consequences of tokening some concept. I again make a working assumption, drawing on the tradition which connects moral judgements to (moralized) reactive emotions. I henceforth assume that for W to play role R is: for the judgement that an agent's action is W makes one blame the agent for so acting (absent a judgment that they have an excuse); and for the judgement that the agent's action is not W inhibits one from blaming them.[2]

In the stability thesis, W ranges over concepts. I take these to be vehicles of mental representation, causally efficacious types which are tokened when we make judgements. I do not need to assume that such types have their conceptual roles essentially (or even that every concept has a conceptual role). The stability thesis, as I understand it here, is silent on whether the tie between the concept of moral wrongness and the wrongness conceptual role is necessary, necessary-

[2] A presupposition must be that *blame,* strictly speaking, is an emotion linked specifically to morally wrongness, and that "epistemic blame", or "blaming oneself" for a prudentially unwise but morally okay course of action are loose speech. If that is denied, then more work would be required in order to narrow down the sentiment to its specifically moral instances. For discussion of this tradition (traced to Mill (1863)), see Darwall (2010) on reactive emotions and moral concepts. Compare Skorupski (2010), p. 292: "It is morally wrong for x to A iff, were x to A from the beliefs that are warranted in x's epistemic state then either x would be blameworthy for A-ing, or extenuating circumstances would apply to x's A-ing" (for Skorupski, blameworthiness is the existence of sufficient reason to feel the sentiment of blame (cf. p. 295)). Skorupski here offers a biconditional analysis, rather than (directly) a description of a conceptual role. There is a connection to (versions of) the conceptual role described, however, for agents who accept the equivalence and feel the way that they judge they have reason to feel. There is a closely related analysis in Gibbard (1990, p. 43): "what a person does is morally wrong if and only if it is rational for him to feel guilty for having done it, and for others to be angry at him for having done it"—except that Gibbard uses "anger" where Skorupski uses blame (and identifies guilt with self-blame).

assuming-rationality, or entirely contingent. It simply asserts the conditional: if a concept plays the relevant role, it picks out wrongness. (Thus a motivational externalist can agree with stability as I have formulated it).

In sum, I'll be working with the following concrete version of the referential stability thesis:

Necessarily, if an agent has some concept W, such that for any act A:

- when the agent judges x's A-ing to be W, this makes them blame x for A-ing (unless they judge that the agent had an excuse);

and

- when the agent judges x's A-ing not to be W, this prevents them blaming x for A-ing;

then

- W picks out the property of *violating the categorical imperative.*

Different views about the distinctive role played by the moral wrongness concept, or about what it is for an act to be morally wrong, will lead to rival versions of the stability thesis.[3] The reader might doubt whether the particular working assumptions I have made are correct. This will not matter. The core argument to be given can be stated schematically, in terms of any role R and property P, and will be valid for any way of filling these in. The reason for making the working assumptions is simply heuristic: a concrete instance will be easier to digest and evaluate. But if, for example, the reader endorses classic hedonistic utiliarianism, and a conceptual role that focuses on first-personal deliberation rather than reactive emotion,[4] they are invited to check that the argument still

[3] I say "rival", but in fact, due to the conditional nature of the thesis, one might endorse multiple stability theses—a variety of conceptual roles, playing any of which is sufficient for denoting moral wrongness.
[4] Wedgwood (2001) offers a first-pass suggestion for a conceptual role for moral wrongness: that the judgement that x is morally wrong leads to the judgement that not doing x is (all things considered) better than doing x. Wedgwood also considers variants of this role which do not make judgements of moral wrongness overriding. Though Skorupski as cited above offers a normative-sentimentalist analysis of wrongness, he argues that moral judgements have an overriding role in practical deliberation, by leaning on a connection between reasons to feel blame and reasons to act (op. cit. pp. 298-299). Scanlon (2007) emphasizes the first-personal deliberative role in his discussions of the concept of moral wrongness, though he emphasizes a "higher-order" role in which (possibly implicit) beliefs about what is right and wrong control how other, more specific considerations figure in deliberation.

goes through on their preferred substitutions. I will flag points at which these differences may matter.

**1.2 Why stability?**

I will presuppose in what follows that stability holds, and that the challenge is to predict and explain that datum (although in section 4.1, I propose a refinement of stability). Foes of stability will need to recast what I establish as a conditional connection between a certain metaphysics of representation and stability, which they *tollens* where others *ponens.* I will not try to defend stability itself here. But to understand the view, it is useful to know the reasons that friends of stability have (or think they have) for endorsing it. Accordingly, I here recap the 'moral twin earth' case that motivates many (Horgan & Timmons 1992; also see Hare 1952. For recent dissent and references to the literature, see Dowell forthcoming).

Suppose the citizens of Utilitas are disposed to blame other agents for taking hedonically suboptimal options. The citizens of Kantopia, on the other hand, blame other agents when they take actions that they cannot will as a general law. We fill in the scenario so each citizen has a concept W ("wrong") that is linked to blame in the way described earlier. The epistemic conditions under which they judge that the concept is instantiated varies by citizenship. Crucially, we are invited to agree that when a citizen of Kantopia tokens "it is W to kill an innocent to save many lives*",* and a representative of Utilitas tokens "it is not W to kill an innocent to save many lives", the two have conflicting views on a common topic. If that is indeed the case, there can be no equivocation in what their respective concepts of wrongness denote: there is some P that they both pick out. This is so even though their views on the extension of W, and so their suggestions as to what property P might be, are diametrically opposed. (On our working assumption, P is the property of violating the categorical imperative, so the Kantopian's moral theory is largely true, while the citizens of Utilitas go wrong.)

A specific “twin earth” thought experiment such as this supports an instance of the universally quantified referential stability thesis, and referential stability in full generality will be supported if analogous verdicts are accepted for other communities (including ourselves) who have a concept W that plays the blame-centric role.

**1.3 Plan**

In this paper I will be presenting a metaphysics of mental representation that predicts and explains stability in the concept *wrongness*. I then argue (*pace* recent literature) that nearby rivals fail this test, and consider objections and replies.

Section 2 of this paper proposes an explanation of the stability thesis. 2.1 introduces radical interpretation; 2.2 articulates the version of it advocated (substantive, realist radical interpretation) and 2.3 applies this to explain referential stability.

Section 3 puts the discussion into the wider context of the theory of content. 3.1 compares the present version of radical interpretation to one often attributed to David Lewis, which has recently been deployed to explain stability. 3.2 presents objections to this rival theory, and shows the present account is not subject to these objections. 3.3 explains how an approximation to the folklore version of Lewis arises on my own favoured view, and uses this to give a common explanation covering both why metaphysically basic descriptive properties (greenness, charge) and normatively central properties (wrongness, personhood) are “normative reference magnets”.

Section 4 consists of replies to objections, leading to refinements of the referential stability thesis itself and clarifying the role that first-order normative theory and normative psychology play in the argument.

## PART II: EXPLAINING STABILITY

### 2.1 A framework for an explanation: radical interpretation

Abstract interpretations of someone—functions that map stages of that person to intentional mental states—are ten a penny. There is one abstract interpretation that maps your current stage to a belief that the Earth is run by the lizard people; another that attributes an overwhelming ambition to count the grains of sand in the Sahara. Most of these abstract interpretations bear no relation to what you actually believe or desire, and make no sense of what you say and do. Still, somewhere amongst them is an abstract interpretation that gets things right.

A metaphysics of mental representation M will give a story in two parts. In the first part, M gives details about this space of abstract interpretations. What kinds of things get interpreted, most fundamentally? Is it whole temporal slices of persons, or does a person at a time have an attitude in virtue of being in a state that is mapped to that contentful attitude? Do our interpretations ascribe flat-out beliefs and desires, degrees of belief and degrees of desire, or both? What attitude-types, other than belief and desire, does an interpretation ascribe? In the second part, M gives an illuminating account of what the world has to be like for one of these abstract interpretations to be correct (and such an account will be reductive if it does not invoke representational states). It is a platitude, more or less, that an abstract interpretation is correct iff the mental states it ascribes are exactly those the agent is in; but this biconditional is not an illuminating or reductive characterization of correctness, so it is not something we can slot into M as an analysis of correctness. However, once we do have a full version of M to hand, we can use this biconditional to convert M's metaphysical analysis of correctness, whatever it may be, into a metaphysical analysis of the facts concerning your mental states: an agent x (at *t*) believes that *p* iff the correct abstract interpretation maps x-at-*t* to a belief that *p*.

The particular metaphysics I will be proposing is a version of *radical interpretation*. The radical interpretation take on correctness, to a first approximation, is the following:

The correct interpretation of an agent x is that one which best accounts for x's dispositions to act in the light of the courses of experience x undergoes.

As I develop the theory, 'actions' and 'experiences' are working primitives in this metaphysical story. I will assume that these notions are understood in representational terms. The possibility of a full reduction of the representational to the non-representational thus depends on the availability of a reductive theory of the content of actions and experiences. What my version of radical interpretation offers is a transformation of a basic kind of "source intentionality" (the proximal ways in which agent and world interact representationally, i.e. actions and experiences) into the "secondary intentionality" of beliefs and desires (see Pautz 2013). With that starting point acknowledged, the core question that distinguishes different flavours of radical interpretation is how we should understand the notion of "best accounting".

Here is a famous and familiar proposal: understand "best accounting for x" as "best structural rationalization of x". The correct interpretation of an agent would attribute beliefs and desires that are rational responses to her experiences on the one hand, and provide means-end rationalizations of her actions on the other. The rational constraints here would be "structural" in that they're insensitive to particular contents of attitudes involved, constraining only the patterns of attitudes and perhaps their logical form. For a concrete example of this, consider (minimal) Bayesianism. At the first stage of M, we take interpretations to map stages of a person to assignments of degrees of belief to each proposition (probability) and degrees of desirability to each proposition (utility). Articulating "best accounting" as "making as structurally rational as possible", we have a Bayesian proposal for the conditions that a correct interpretation should meet:

- Rationality constraints on beliefs over time: that they are updated by conditionalization on experience.

- Rationality constraints on beliefs at a time: that they are probabilistic (i.e. satisfy the axioms of probability theory).
- Rationality constraints on final and instrumental desires: that they fit means-end constraints articulated by a formal theory of decision (such as that of Jeffrey 1967).
- Rationality constraints on choices: that the agent chooses to do the thing they most desire to do, among the things they think they're able to do.

This gives us a candidate metaphysics of representation: *structural* radical interpretation (Bayesian-style). An interpretation will be correct if and only if it comes closest to making the agent's dispositions to act (at a time *t*) dispositions to select the option that maximizes expected desirability (according to the assignment at *t*), and to making the beliefs and desires attributed evolve under the impact of experiences in the way the Bayesian demands.

Structural radical interpretation as it stands does not work: perhaps the correct interpretation meets all its constraints, but David Lewis (1983) demonstrated that wildly inaccurate interpretations do so too. I discuss this in detail elsewhere (Williams 2016), but for present purposes, it suffices to picture Lewis's argument as a black box, which takes as input an arbitrary sensible interpretation and produces as output a deviant variation on that interpretation. The two interpretations coincide on the agent's attitudes concerning what goes on in a local space-time bubble surrounding her, but the deviant one represents the agents as agnostic and indifferent to matters outside that bubble. If we stick to structural rationalization of actions in the light of experiences, it turns out we cannot eliminate deviant bubble-interpretations of agents. Something more is needed.

**2.2 Substantive rational interpretation**

Bayesian constraints do not rule out wild initial belief states. Given some such set of wild initial beliefs, it will be rational by Bayesian standards to have high conditional confidence that the world will explode tomorrow, given the (entirely ordinary) course of experience you have undergone to

this point—and so rational, after undergoing that experience, to end up believing that the world will explode tomorrow. Nor do Bayesian constraints rule out wild final desires, e.g. basic desires for a saucer of mud or indifference to what happens to you on future Tuesdays. That is by design: the core Bayesian story as articulated above was developed as a theory of the formal patterns that a well-run mind should exhibit, not about the particular contents that we have most reason to believe or desire. It is, after all, a theory of structural rationality.

Yet there's something crazy about a basic desire for a saucer of mud, or future-Tuesday indifference, and about humdrum experience triggering paranoid beliefs—something deeper and more alien than what's wrong with commonplace false beliefs or unwholesome desires. In addition to constraints of rationality based on formal patterns among our attitudes, perhaps there are rational constraints that are sensitive to the particular contents we think or want. These would be constraints of substantive rationality.

Substantive radical interpretation is accordingly a metaphysics of content on which the correct interpretation of an agent is the one that does the best job of making her substantively as well as structurally rational. This is what Lewis (1974, 1992) advocates. Unfortunately, Lewis never told us much about what these constraints of substantive rationality were, beyond giving a few examples. What metaphysics of representation we get out will depend on what account of substantive rationality we feed in. To make progress, we need more information.

The position of this paper is that substantive rationality is a matter of an agent's *reason-responsiveness* — a measure of the extent to which the agent is responding as they should to good normative reasons for their beliefs and action. This can account for the substantive-rationality data. An action motivated by a basic desire for a saucer of mud is not performed for good normative reason (good normative reasons for the act may exist, but they are not the agent's reasons, ex hypothesi). There is, I take it, no good normative reason for future-Tuesday indifference. Likewise, there is no good normative reason to believe that the world will explode, if

one's course of experience is entirely humdrum. This reason-responsive gloss on "substantive rationality" links radical interpretation to first-order epistemology and first-order theory of action. In so doing, it makes substantive radical interpretation predictive, in a way it would not be if we left "substantive rationality" as a mere placeholder. Wider theory gives us a fix on (epistemic or practical) normative reasons, and what it is to respond to these reasons appropriately. From this, the metaphysician of representation reads off consequences for which interpretations of agents best account for their dispositions, and thus (given substantive radical interpretation) for which interpretations count as correct.

The way in which substantive radical interpretation (on this reading) favours some structurally rational interpretations over others can be illustrated by epistemic inductive reasons (compare Pautz 2013, Schwarz 2014, Weatherson 2013). Suppose you have encountered a sample of five hundred emeralds after a systematic search, all of which have been green. Post-Goodman (1954), we know that structural rationality alone allows many different doxastic responses. With appropriate priors, one could recognize all that evidence, and assign high confidence to *all emeralds being green* or instead to *all emeralds being grue* (i.e. green and first observed before T, or blue and first observed after T). Either response can be made perfectly structurally rational by assuming the right background beliefs. But normative epistemology may still distinguish the cases. According to one such epistemology (call it E), these observational facts give you reason to believe that all emeralds are green; they do not give you reason to believe that all emeralds are grue. If E is correct, then all else equal, an interpretation will depict an agent as *more reason-responsive* if it represents her as a green-inductor, than if it represents her as a grue-inductor. Substantive radical interpretation would then favour the former over the latter (all else equal).

The green/grue example illustrates two aspects of substantive radical interpretation. First, it shows how a particular theory of epistemic reasons can generate new constraints on interpretation. We'll see the pattern repeated below, but in the practical rather than epistemic sphere. Second, the inductive case highlights how the normative premises required are not truisms, but can themselves

be matter of first-order dispute. A subjective Bayesian should reject the claim that a grue-inductor is any less reason-responsive than a green-inductor. The premises about practical normativity below will be contestable in the same way.

**2.3 Resolution of the stability puzzle**

The underlying strategy for explaining the referential stability of wrongness is very simple: only by interpreting W as wrongness do we make the patterns that are built into the conceptual role it plays reason-responsive. Combined with substantive radical interpretation, this predicts and explains the fact that W denotes *wrongness.*

Let us work this through for a particular case. Sally has learned that Harry has cheated on his exams, and judges that Harry's cheating is W. She has also learned that Harry engaged in tactical voting, and judges that this is not W. What property does her tokening of W denote? Assuming that Sally's tokenings of W play the blame-centric conceptual role, I'll argue that substantive radical interpretation entails that W denotes *wrong.*

First, W plays the wrongness-role in Sally's cognitive economy: Sally's judgement that Harry's cheating is W, for example, makes her blame him for that act (since she doesn't take him to have an excuse). Moreover, her judgement that Harry's tactical voting is not W makes her not blame him for tactical voting.

Second, according to substantive radical interpretation, the correct interpretation is the one that makes Sally maximally reason-responsive, and so in particular must make the way she handles [A is W] judgements maximally reason-responsive. Additionally, I assume what I will call "conceptual-role determinism for wrongness": the interpretation on which Sally is most reason-responsive overall is one on which her concept W (which plays the blame-centric role R) denotes something that makes the link between her W-judgements and blame most reason-responsive. Absent this assumption, it might be that although a given interpretation does best at making the links between

W-judgements and blame most reasonable, this is outweighed by other countervailing factors; essentially it is the assumption that "all else is equal". I discuss the assumption in section 4.1 but for now I ask the reader to just take it on board.[5] Putting substantive radical interpretation and conceptual role determinism together, we derive the following: the correct interpretation of Sally is one that makes the link between her W-judgements and blame most reason-responsive.

The third and final element in the explanation is are normative premises that tell us what it takes to make the right kind of sense of the blame-link:

1. A reason-responsive agent would be such that judging that Harry's cheating was wrong and unexcused makes them blame Harry for cheating (directly and indefeasibly).
2. A reason-responsive agent would be such that for any feature F other than those that entail wrongness, the judgement that Harry's cheating was F (and unexcused) would not make them blame Harry for cheating (directly and indefeasibly).
3. A reason-responsive agent would be such that: the judgement that Harry's tactical voting was not wrong, would make them refrain from blaming Harry for tactical voting (directly and indefeasibly).
4. A reason-responsive agent would be such that: for any feature G other than those that wrongness entails, the judgement that Harry's tactical voting was not-G would not make them refrain from blaming Harry for tactical voting (directly and indefeasibly).

[5] Some theorists think of conceptual roles as psychological natural kinds, and assume that any concept whatsoever comes equipped with its own specific conceptual role. That is not at all the way I am inclined to use the term. For me, conceptual roles are no more than a patterns in the way we deploy a concept with a certain theoretical interest---in the present context, they label the patterns used to articulate the stability thesis. I do not assume there is any topic-neutral way of picking out a conceptual role, and so I am not inclined to see the conceptual role determinism thesis as an instance of a more general thesis connecting "conceptual roles" to the theory of reasons. If one were more of a realist about conceptual roles than I (cf Peacocke) then you might regard it this way, and I am not opposed to that reading. Just to put a different view on the table to temper that way of thinking: one might think of conceptual role determinism as resulting from the following stipulation: let us call nothing a conceptual role unless it relates to the theory of reasons so as to make conceptual role determinism true (given that stipulation, the key question is whether the blame link *is* a conceptual role in the first place). For now, I simply rely on the condition, and discuss its status later.

By (1), an interpretation of Sally on which W picks out *moral wrongness* (= violating the categorical imperative) will make her W-involving dispositions reason-responsive. By (2), only interpretations which denote features which entail moral wrongness can do the job. In particular, this rules out interpreting Sally's concept W as denoting *failing to maximize hedons,* since ex hypothesi such an interpretation will not make Sally reason-responsive. Notice: this is so even if Sally herself were a convinced utilitarian, and so mistakenly thinks that to be W is to fail to maximize hedons.

This doesn't quite get us to the conclusion that W denotes *moral wrongness*. For all that (1) and (2) tell us, we might interpret Sally's W as denoting something more specific than moral wrongness which nevertheless entails it—perhaps the property of *being deceptive in pursuit of self-interest.* That is why we need (3) and (4), which tell us what kinds of properties would make sense of the link between Sally's negative W-judgements (that she thinks that it is not W to vote tactically in elections) and the fact that this blocks her blaming Harry for his avowed tactical voting. The pattern of links between Sally's negative W-judgements and blame is not reasonable under the over-specific interpretation of W as *deceptiveness in pursuit of self-interest*, since on that interpretation Sally wouldn't have excluded the act as blameworthy in some other way. The interpretation of W as wrongness, however, does make sense of this second aspect of the conceptual role of Sally's W.

In sum: from these four normative premises, together with substantive radical interpretation and conceptual role determinism, we reach the conclusion that Sally's concept W denotes moral wrongness.

To derive referential stability in full generality, we need to strip away any idiosyncrasies of Sally's case. The four premises above concern specific acts of cheating and tactical voting. In the general case, we need a generalized form, for example in the case of premise 1:

> A reason-responsive agent would be such that: for any agent x and act A, the judgement that x's A-ing was wrong (and unexcused), makes them blame x for A-ing.

Mutatis mutandis for premises (2-4). The form of the argument would be the same, so it will be valid just in case the instance above is. I discuss the soundness of the generalized form further in section 4.3.

Another step of abstraction can illustrate something I promised earlier: that the explanation for stability I offer does not depend essentially on the working hypothesis about the conceptual role of wrongness, or the identity of moral wrongness itself. This is because no matter what how you fill this in, we can run the same form of argument by appropriately filling in the following schemata for normative premises:

> A reason-responsive agent would be such that: for any x and A, as a direct and indefeasible result of a judgement that x's A-ing is wrong, they will enact conceptual role R.

> A reason-responsive agent would be such that: for any G other than wrongness, and for any x and A, as a direct and indefeasible result of a judgement that x's A-ing is G, they will not enact conceptual role R.

The reader is invited to try this out for their favoured conceptual roles/identifications of wrongness.[6] The content of the normative premises that is needed will of course vary, and as flagged earlier,

[6] For example, theorists like Scanlon (2007) and Wedgwood (2001) link the conceptual role of wrongness to deliberation and preference over how to act, rather than reactive emotions. Wedgwood's initial suggestion on this front is that moral wrongness has a distinctive overriding role in fixing preferences about x. If we had adopted this as our working assumption, then the version of 1 under consideration would state that a reason-responsive agent's judgement that Harry's cheating was wrong would make them prefer Harry not cheating over Harry cheating.
You might also think that there is no privileged conceptual role for moral wrongness—perhaps both the Wedgwood-style role in practical deliberation and the Skorupski-style role in reactive emotion are enacted in different communities, different people, or different stages of one person. So long as the appropriate

the soundness of the derivation will depend on the first-order issue of whether the normative premises that result obtain.

I finish this section with two clarificatory comments, with worries deferred to section 4.

First, the explanation I offer is consistent with the obvious truth that the causal history of a particular episode of blame can involve all sorts of property-ascriptions. Suppose again that Sally is a convinced utilitarian, and thinks (ex hypothesi incorrectly) that to be morally wrong is to fail to maximize hedons. Sally judges that Harry's cheating fails to maximize hedons, and this ultimately leads her to blame Harry for cheating. The path from judgement to blame here is indirect: she will infer from her hedonic belief and her background moral views to the conclusion that Harry's cheating is wrong, which is what most immediately makes her blame him for cheating. This is why the premises have the "direct and indefeasible" rider. This is crucial to the plausibility of the uniqueness claims in premises (2) and (4).

Second, a structural point: the normative premises are not conditionalized to agents with suitable aims, tastes or preferences---it is assumed that any reason-responsive agent whatsoever will satisfy them. Theorists such as Railton (1986) and Boyd (1979, 1982, 1988) would reject the implicit assumption that (moral) reasons in question are categorical in this way. They argue instead that moral reasons are conditional (albeit conditional on features that are deeply entrenched in human nature). My argument presupposes that Railton and Boyd are mistaken on this point. This is not merely a working assumption (as with my identifications of R and P earlier) but essential to the argument as stated. However, if Boyd and Railton are correct and moral reasons are hypothetical rather than categorical, an adjusted version of the argument will nevertheless go

analogues of (1) and (2) both hold, we can argue that the diversity of roles converge on the single property of moral wrongness. Skorupski, for example, argues that an overriding role for moral judgement in practical deliberation follows from his blame-centric account of the concept of wrongness, together with conceptual truths about the way blame relates to reasons. This suggests that the argument can be further abstracted, to argue for something broader even than referential stability. I will not explore this further here however.

through. One would alter (1) and (2) by making them conditional on the agent having aim T. One would derive a restricted form of stability: for all agents with aim T, and a concept W that plays role R, W denotes P. This is not the full, unconditional stability that is my target, but some might prefer the package deal of hypothetical moral reasons and restricted stability. It still allows agents with utterly divergent views of the extension of W to denote *moral wrongness*, since sharing aim T is prima facie independent of whether one subscribes to Kantianism, Utiliarianism or something else entirely. I now set this aside and persist with my original formulations.

## PART III: AGAINST REFERENCE MAGNETIC ALTERNATIVES

Section II presented a version of radical interpretation and argued (given auxiliary first-order normative assumptions) that it predicts and explains the referential stability of moral wrongness. I believe this to be an elaboration of the metaphysics of mental content that David Lewis endorsed. In the recent literature, a Lewisian metaphysics of content which gives pride of place to the notion of certain properties being "reference magnets", has been offered in explanation of referential stability. This section contextualizes my explanation of stability by examining this alternative approach, highlighting its problems, and showing how the view endorsed here relates to it.

In 3.1 I present an alternative ("folklore") version of Lewisian interpretationism. Section 3.2 presents Dunaway and McPherson's case that this theory explains referential stability. I argue that it does not. Section 3.3 shows how the folklore Lewisian idea that certain metaphysically basic properties are "reference magnets" can be derived within substantive radical interpretation. I argue that this should be seen as a special case of a more general notion of a normative reference magnet which is operative in my own explanation.

### 3.1 Folklore Lewisianism

Lewis offered separate accounts of the metaphysics of mental and linguistic content, and the folklore version of his interpretationism is based on his remarks on the latter. The canonical source of this version is Lewis (1983)'s reconstruction and response to Putnam (1980, 1981). We start by

supposing that we can identify circumstances in which agents are assenting to certain sentences in their public language. From this, we attempt to extract a certain "folk theory" the agent endorses. We will suppose that this consists of those sentences she is disposed to assent to come what may—that she treats as platitudes. The successful interpretation must at least render the agent's folk theory true. Variations on the theme are possible: for example, we might replace the appeal to an individual's folk-theory with that of the folk theory implicit in a whole linguistic community.

The constraint of making folk theory true wildly underdetermines reference (that was one of Putnam's points). Some additional metasemantic factor needs to be introduced to correct for this (that was Lewis's). At this point, the distinctive gambit of folklore-Lewis enters: a metasemantic role for metaphysically basic properties. Some properties (maybe the properties featuring in final physics) are metaphysically basic or "perfectly natural". Other candidate semantic values are graded as more or less natural, depending on how close they are to the perfectly natural, and how to understand "closeness" here is another locus of variation. An overall interpretation of a language is graded as more or less *eligible* on the basis of the overall naturalness of the semantic values it assigns to expressions.

So the metaphysics of linguistic representation in question posits a way to rank interpretations of language by their *degree of eligibility* (a measure of the overall naturalness of semantic values assigned) and *degree of fit with usage* (a measure of how much of folk theory they make true). The correct interpretation is the one that optimizes both eligibility and fit.

The introduction of eligibility makes fundamental/natural properties "reference magnets". All else equal, you end up thinking and talking about more eligible properties rather than their less eligible rivals, even when the way you use the language could be construed as thinking or talking about either with equal charity. This reference magnetism is absolute, a factor that matters to reference-fixing that is entirely independent of the way we use the language.

What has just been sketched is a metaphysics of linguistic content, for public language. It has a very different form from substantive radical interpretation, which is a metaphysics of individual mental content. But it is possible to adapt the above ideas into an "interpretationist" account of mental content that would be a rival to radical interpretation. Suppose (following e.g. Field 1978) that what it is to believe that p is to have a mentalese sentence S in one's belief-box, and for that sentence to say that p. Interpretations of the agent can then take the form of interpretations assigning truth-conditions to sentences of the agent's mentalese language. An agent's folk-theory could be identified, not with a class of public language sentences she will endorse come what may, but a class of mentalese sentences that she is disposed to token-in-her-belief-box come what may. There is a coherent eligibility-invoking metaphysics of mental content to be given along folklore Lewisian lines, and one way we can evaluate it is by looking at its consistency with referential stability.

**3.2 The folklore and referential stability**

According to a folklore Lewisian, if a cluster of properties are in the running to be reference of "F", and one is much closer to the fundamental than the others, then that will end up as the reference—and robustly so, since modest variations in how well that interpretation fits with usage will be swamped by the big differences in eligibility.

Dunaway and McPherson (2016) build an account of the referential stability of normative words based on eligibility. Let's suppose the Kantopians are right about what permissibility is—it requires one adhere to the categorical imperative. These authors argue that the property *coherence with the categorical imperative* will then be highly natural, and so interpretations ascribing it as reference of a predicate will receive a boost to their eligibility score.

Turning to the populace of Utilitas, more of their folk theory would be made true by assigning *maximizing hedons* as the property picked out by their term "permissible", but this charity-boost will be swamped (argue Dunaway and McPherson) by the loss of eligibility incurred. Overall they say

that even for that population, the interpretation on which "permissibility" picks out *coherence with the categorical imperative* scores best overall when charity and eligibility are traded off.

In order to make this case, Dunaway and McPherson need to justify the claim that permissibility itself (whatever it turns out to be) is a lot more eligible than rival properties. That looks like a tall order if relative eligibility is determined by definitional distance from the properties of microphysics, which is the gloss on "closeness to the fundamental properties" that Lewis seems to endorse. Accordingly, they drop this aspect of the Lewisian package. Instead, they propose that the basic properties of any serious science count as perfectly natural. Serious science need not be microphysics! Indeed, they claim, moral theory itself has this status, and its basic properties (in particular, permissibility) thereby count as eligible to the maximal degree.

Schroeter and Schroeter (2013, p.20), anticipating and exploring this sort of response, raise a worry: "if conventional morality, preference maximization, maximization of human flourishing, and conforming with the categorical imperative are all basic joints in nature, then being a joint of nature won't help". Their point is that if eligible properties abound, the eligibility of moral permissibility will not be a very significant metasemantic filter. The first thing that Dunaway and McPherson might do is resist the antecedent of the Schroeter and Schroeter conditional—they claim eligible properties must feature in a serious science, and there's no reason they need to grant that e.g. conventional morality plays this role, if it does not pick out moral permissibility. But ultimately this style of concern is fatal to the Dunaway and McPherson story. Schematically: suppose that E is a community who have a concept that plays the internal conceptual role of *wrongness,* and who take this to pick out a property P where (perhaps unbeknownst to them) P plays a basic role in some *other* serious science. Perhaps, for example, *maximizing self-interest* is a property that has a basic explanatory role elsewhere in normative theory (or economic theory, or psychological theory…); and E are a community of egoists, thinking that the right action (for x) is that which maximizes x's self-interest. In virtue of its explanatory role outside morality, the property onto which E have latched will be maximally eligible---just as eligible as authentic moral permissibility. In virtue of their

egoist moral theory, interpreting "morally wrong" as *failing to maximize self-interest* will score better than rivals in terms of maximizing truths attributed, and is no worse on the dimension of eligibility. So here we have a case where a community has a concept that plays the internal role of moral wrongness, but fails to thereby denote moral wrongness itself. That is exactly what we needed to avoid.

This is a structural defect, not a puncture to be patched. Whatever makes for eligibility, so long as we can find a maximally eligible property of actions that differs from moral wrongness, then we can envisage a community who mistakenly take it to be the property that constitutes moral wrongness. And it is hard to envisage a story about what makes for eligibility that allows in the normative good guys (moral or all-things-considered wrongness) but rules out false friends from natural or social sciences or normative theory (being caused by chemical agents ABC, selection by certain frugal-but-faulty human heuristics, prudentially wrong).

Dunaway and McPherson's intention was to show that eligibility-laced Lewisian metasemantics can explain referential stability. I've argued they fail. En passant, this establishes something very interesting: a novel reductio of the influential folklore Lewisian metasemantics. (This presumes that they've made the best case possible for it within that framework, and I think they have.) [7]

**3.3 Deriving the magnetism of the natural**

What role might metaphysically eligible properties have from the perspective of substantive radical interpretation? Pautz 2013 and Weatherson 2013 give us the needed bridge.

Let us go back to the case of induction introduced in section 2.3. Suppose that an agent infers that all emeralds are G from past observations of emeralds which were each G. Here are two candidate interpretations of the agent: the first interpretation says that G = green; a second interpretation

[7] This isn't to rule out modifications that significantly adapt the folklore to avoid this worry---see the van Roojen below, for example.

says that G = grue. Both interpretations have the agent inferring according to the formal pattern of enumerative induction.

Reason-maximizing substantive radical interpretation will rate the former interpretation higher than the latter if there is normative epistemic reason to believe that all emeralds are green, and no epistemic reason to believe that all emeralds are grue in the circumstances described (all else equal). It will do so, no matter what the agent's other beliefs are, so long as these facts about normative epistemic reason obtain irrespective of background beliefs. In sum: given suitable auxiliary epistemological assumptions, similar in kind to those we needed to explain referential stability of wrongness, projectable properties are reference magnetic.

Now add the further epistemological assumption that properties are more projectable the more metaphysically elite they are—e.g. that there is greater *pro tanto* epistemic reason to inductively generalize on P compared to Q, to the extent that P is more natural than Q (compare Weatherson 2013). Call this (contentious!) bit of epistemological theory Inductive Metaphysical Privilege. Adding this to the mix allows us to argue that, with respect to the interpretation of concepts on which we induct, metaphysical natural properties are reference magnetic. This is the suggested reconstruction of the magnetism of metaphysically eligible properties.

Suppose we accept this as a reconstruction of a reference-magnetic role for natural properties within substantive radical interpretation. And to be sure, we induct on "moral wrongness", and so for a reason-responsive agent its referent must be projectable. But lots of properties are projectable (a datum that any precisification of Inductive Metaphysical Privilege is going to need to save, in order for it to be credible). But moral wrongness (=violating the categorical imperative) is projectable, as is failing to maximize hedons and failing to maximize self-interest. The constraint of making the agent's inductive practice reasonable may make all of these "more magnetic" than gerrymandered rivals, but (again on credible precisfications) it won't by itself favour one of the projectable candidates over others. That means the reconstructed role for natural properties within

substantive radical interpretation can't by itself explain stability – but by the same token it avoids the refutation of the folklore Lewisian account set out above.

To summarize the morals of this section: pace Dunaway and McPherson, one cannot use reference magnetism and the folklore Lewisian metaphysics of reference to explain referential stability. One can, however, use substantive radical interpretation (with appropriate normative side-premises, such as Inductive Metaphysical Privilege) to explain why the folklore Lewisian story was appealing in application to the kind of descriptive concepts to which it is typically applied. There are significant differences. On the relocated account, metaphysically eligible properties will only be "reference magnets" relative to a certain feature of the conceptual role of the target concepts—that we are prepared to deploy them in induction. Similarly, in the explanation that I offered above, it's only because of the conceptual role of Sally's W that the property of moral wrongness (=violating the categorical imperative) becomes an especially magnetic referent.

Incorporating relativity into a Lewisian metasemantics has been proposed before: Van Roojen writes: "The kinds or properties which are more natural for the purposes of physics may not be the same as those which are more natural for purposes of biology. The more eligible semantic values for one's terms when engaged in the former may or may not be the same as the more eligible semantic values for one's terms when one is engaged in the latter." (van Roojen 2006). According to van Roojen, there are D-eligible properties that are more eligible for thoughts one has when one is engaged in discipline D. This sort of relativization is problematic, however. As Schroeter and Schroeter note, *engagement in a discipline* is prima facie an intentional specification of a stretch of thought or talk, and this makes it hard to see how van Roojen's particular brand of relative eligibility can be deployed within a reductive metaphysics of representation.

The relativity in play according to substantive radical interpretation is of a different sort: it is relativity to *role in one's conceptual economy.* Consider the explanation of the last section: the agent has to be deploying a concept in the right way (e.g. inductively generalizing on it) for

considerations of inductive reason to be relevant at all. There is no general bias towards attributing projectable properties as the interpretation of concepts in substantive radical interpretation, as there is for the folklore Lewisian. I have the concept “grue”, but I am not at all inclined to deploy it in inductive reasoning; an interpretation which assigns that concept a projectable property is not favoured, indeed, it would be disfavoured, for if “grue” referred to a projectable property then by refusing to induct upon it I would be out of line with my epistemic reasons. In the same way, the particular blame-centric conceptual role that we assumed was characteristic of the concept of moral wrongness was central to the explanation of referential stability in section 2.

While referential stability amounts to a counterexample to folklore Lewisian theory of reference magnets, according to substantive radical interpretation that account remains a useful model for one factor relevant to reference determination for a (descriptive, inductive) fragment of our thought and talk. By the same token, it is very misleading if taken as a general model for concepts irrespective of their conceptual roles. In the first instance, substantive radical interpretation gives us *normative* reference magnets (and always relative to conceptual role), drawn from the properties that populate the theory of reasons. It is only by relying on a thesis about the character of inductive reasons that we get metaphysically elite properties into the theory at all. Relative to appropriate conceptual roles, *wrongness* has just as a much a claim to be a reference magnet as *greenness* or *charge*.

## PART IV: REFINING THE STABILITY THESIS

In the remainder of the paper, I will be considering objections to my favoured explanation of referential stability, using them as opportunities to clarify the account.

### 4.1 Conceptual role determinism

Conceptual role determinism is the thesis that the most *overall* reason-responsive interpretation of x’s concept of moral wrongness W is one which makes certain local and specific aspects of their mental economy (the link between W-judgements and blame) maximally reason-responsive. This

was the assumption that allowed an interpreter to ignore possible trade offs between different ways that W is deployed in an agent's mental economy, and concentrate on the conceptual role alone.

Some may worry: "Conceptual role determinism was a premise in the explanation of stability, but we haven't yet heard anything to explain why it should be true, in general. And if it isn't, then we are not entitled to claim that referential stability has been established in full generality" . Call the following the Refined Stability Thesis:

> Necessarily if an agent has a concept W that plays role R *and conceptual role determinism holds* for W, then W denotes P.

The objector is pointing out that original stability thesis did not include the italicized restriction, and so what the argument establishes is not what was originally advertised. I am going to concede the point, but argue that the Refined (not fully general) version of stability is what we should have been targeting all along.

First, a case where Conceptual Role Determinism plausibly fails. Consider Suzy, who has a concept W that plays the blame-centric conceptual role. However, Suzy in addition has a basic indefeasible disposition to infer *act A is W* from *act A does not maximizing*.

No single interpretation of W can make Suzy entirely -- practically and epistemically -- reasonable. Interpret her concept W as *moral wrongness* (=*violating the categorical imperative*) will make her W-judgment forming disposition unreasonable. Interpret her W as *failing to maximize hedons* will make unreasonable the way her W-judgments link to blame. Suzy deploys W in multiple ways, and these ways are in tension with one another. We're not entitled to Conceptual Role Determinism for Suzy's W—that would be to assume that it is more important (for maximizing overall reasonableness) to make sense of the way W relates to blame than the way that W-

judgements are formed. I see no principled basis for that assumption. Henceforth, I will take it that (suitably elaborated) Suzy provides a case where Conceptual Role Determinism fails.

Suzy has an extremely strange (almost tonk-like) moral concept, in which she treats a person's failure to maximize hedons as *analytically* grounds for blaming them. More ordinary moral error does not afford exceptions to Conceptual Role Determinism. Since ordinary moral error is common, and it will be important that exceptions to Conceptual Role Determinism are rare, I pause to examine these cases.

For an illustrative example, consider our interpretee Sally. To be wrong (we are assuming) is to violate the categorical imperative, but Sally mistakenly thinks to be wrong is to fail to maximize hedons. However, Sally was responsible in reaching this moral belief: she considered a range of actual and possible cases, and consulted and argued the issue through with trusted kith and kin. Ultimately, weighing her evidence, she endorsed utilitarianism. The contrast with the way that Suzy formed her W-beliefs is stark! I say:

(i) Sally has the reasonable (though mistaken) theoretical moral belief that to be wrong is to fail to maximize hedons.

(ii) Sally has a reasonable direct and indefeasible disposition to blame those who (she judges) act wrongly without excuse.

(iii) Given (i) and (ii), Sally has a reasonable (if mistaken) derived disposition to blame those who (she judges) fail to maximize hedons without excuse.

The first claim is simply an observation about Sally's epistemic practice. The second summarizes the normative assumptions at the heart of the earlier argument. Having a (derived, defeasible) disposition to blame those she judges fail to maximize hedons is inevitable given (i) and (ii) and a modicum of rationality. (iii) is surely reasonable, if the underlying (i) and (ii) are.

So I suggest that there's no reason to doubt that conceptual role determinism holds in Sally's case. To be sure, by reinterpreting her moral concept W so it denoted *failure to maximize hedons*, we

would avoid attributing false beliefs---under that interpretation, she would have reasonable *and accurate* W-beliefs. But the increase in accuracy would come at the cost of making the basic blame-dispositions in (ii) unreasonable. In Suzy's case at this point, we were faced with a nasty trade-off between epistemic and practical reasonability. In cases of ordinary moral errors like Sally's we have a tradeoff between reasonability on the one hand, and accuracy on the other – and that sort of trade-off simply doesn't threaten Conceptual Role Determinism.[8]

(The contrast between Sally and Suzy highlights a way in which the exact details of the agent's mental structures are central to my explanation. Both Sally and Suzy are disposed to blame someone, when they judge that that person has failed to maximize hedons without excuse. In Suzy's case, in effect I say that there's no way of interpreting this so that Suzy is reasonable, since that is (ex hypothesi) not a good reason for blaming someone, and it's a *basic* feature of Suzy's mental economy that she treats it like it is. But exactly the same pattern in Sally is perfectly reasonable, I say, since the disposition is not basic, but derived from two individually reasonable states.) [9]

Let me turn now to the significance of (rare) exceptions to Conceptual Role Determinism. The first thing to note is that cases like Suzy's show more than that I have failed to derive the original, unrestricted, referential stability thesis. If Suzy's case is possible, then it provides a counterexample to the (determinate) truth of that original thesis. After all, Substantive Radical Interpretation applied to Suzy says that the correct interpretation is the one that makes her most reasonable. But we said above that there it was not determinately the case that interpreting her W

[8] The same thing goes for less theoretically opinionated agents who have no overarching theory that structures their moral thinking---the analogues of (i) for an ordinary untheoretical agent will be a range of hard-to-systematize context specific judgements of the form: Harry judges that what Z did was wrong on grounds it was x, y, z. The analogue of (iii) will be a derived disposition to blame Z for so-acting, on the basis that it was was x, y, z. But each instance of (i) is reasonable, then again the induced instances of (iii) will be reasonable, and there will be no interpretative pressure to set against the need to make reasonable the blame-links in (ii).

[9] A reader for this journal noted (correctly!) that this opens up a couple of ways to cause trouble for my account: either to argue that such nuances (between basic and derived dispositions) cannot make a normative difference; or to argue that the moral twin earth intuitions that motivate stability are insensitive to such differences, so they cannot play the crucial role I'm attributing to them.

as *moral wrongness* made her most reasonable. So despite the fact that W plays the blame-centric conceptual role in Suzy, it does not determinately denote moral wrongness.

I contend, however, that this is an independently plausible thing to say about Suzy's case, a datum that a successful theory of reference should respect. So the failure to derive an unrestricted referential stability thesis is an excellent thing! The cases that motivated referential stability (and were problems for rival accounts) are cases such as Sally (an agent subject to ordinary moral error), or the citizens of Utilitas and Kantopia. In all these motivating cases, say I, conceptual role determinism holds. The motivation that friends of referential stability had for their universally quantified claim was simply that such parade cases are representative of the general case. But the generalization is not obviously warranted, and we have seen that cases like Suzy's show it to be flawed. It is reasonable methodology to use a metaphysics of reference as a tool to predict and explain the scope of the generalization. It is in this spirit that I offer the Refined Stability Thesis.

(There is an alternative strategy the interested reader may wish to explore: to add content to the conceptual role R itself, over and above the links to blame, from which conceptual role determinism can be derived. Suzy's W, we would then say, does not play the relevant role R, and so she is outside the scope of the thesis anyway. If one found a way to do this, it would elegantly deal with these problem cases. (Compare inferentialist harmony constraints and Suzy's case to that of the unharmonious "logical concept" *tonk*). The strategy is independently interesting, but I see no reason to prefer it to the simple refinement of stability above---after all, if conceptual role determinism followed from R itself, then the original and refined versions of stability would be equivalent!)

**4.2 Reasons and reason-responsiveness.**

Some may worry: **"**The normative premises in the "derivation" of referential stability are highly contentious. They tell us that Sally is being reason-responsive when she reacts to a judgement that Harry's cheating is *wrong* by blaming Harry for cheating. But that is a misdescription! The reason

that Sally has for blaming Harry for cheating is not the thin property of *wrongness*. It is the more specific *wrongmaking features* of his act—it's deceitfulness, say."

The objector may elaborate this in various ways. To think that wrongness is a reason (for feelings of blame, as much as action) is to "double count" reasons (Dancy 2000, 2004). Or perhaps "Someone who is moved primarily by overall moral verdicts, such as an action's being wrong, rather than the individual grounds for such verdicts, has been said to have a kind of moral fetish, rather than being truly morally good" (Darwall 2010, p. 136, alluding to Smith 1994). Finally, one might worry that the kind of normative premises I rely on encourage a view of moral psychology where there must be "one thought too many" (Williams 1981)—where the move from the classification of Harry's cheating as deceitful to blame must await the classificatory thought that deceit is wrong.

My derivation is innocent of such charges. To start with the "one thought too many" worry: my premises are entirely consistent with thick moral judgments prompting blame. Sandra's judgement that Harry's cheating is deceitful can make Sandra blame Harry for cheating, in exactly the same direct way that a judgement of wrongness would do (the uniqueness premise I relied on was that only features that entail wrongness can directly make a reason-responsive agent feel blame, which of course allows for Sandra's case). Nor can I see anything in the account that gives primacy to the thin over the thick in this respect. Some think that the right analysis of thick moral concepts already involves the thin concept of wrongness (Elstein and Hurka 2009), and given that view, one might be tempted to defend the explanatory primacy of a wrongness-blame conception. But that's no commitment of mine.

Turning to the other concerns, I am also neutral over whether *being wrong* is itself a *reason* to blame the agent. The normative premises that I used in my explanation (like substantive radical interpretation itself) are formulated in terms of what a reason-responsive agent would feel/do, and not what normative reasons that agent has for feeling/acting that way. I have not taken a stance

here on the relation between reason-responsiveness and the theory of reasons. Dancy (2000, 2004) and anti-Dancians (Heuer 2010, Darwall 2010) will take very different views of this question. It is open to anti-Dancians to identify a reason-responsive agent with an agent who believes/acts/feels for normatively good reasons. Dancians however think of normative reasons as always sparse and specific, and a consequence will be that we'll often know that there are reasons to feel a certain way, without being able to identify what those reasons are.

Suppose Sally knows that Harry has tattled, and knows that this was either spiteful or deceitful, but lacks the background context that would let her determine which it is. The thing she does know (that it was either-spiteful-or-deceitful) is not itself a normative reason for anything, ex hypothesi. Any plausible Dancian theory of reasons will need to accommodate the fact that in Sally's disjunctive-information situation, she should blame Harry for tattling, despite her inability to identify a (sparse/specific) reason to blame him. It will need to theorize what it is for an agent to be "reason-responsive" in cases of limited information. The case of wrongness is analogous. If Harry has spitefully tattled, but all Sally knows is that Harry's tattling was wrong (which entails that there is reason for her to feel blame), as a reason-responsive agent, she should blame him for tattling.

**Section 4.3 Interpreting the morally depraved**

A final worry: **"**Your explanation of stability is all very well in cases like Sally's, or to sensible Kantians or Utilitarians. After all, we take such theories seriously, so it's no accident that we are happy to class adherents as responding to reasons when they blame someone. We expect them to blame people who are worthy of blame a lot of the time, diverging only in recherché cases. But what of agents in the grip of an utterly depraved moral theory—valorizing personal honour and blaming others who indulge in acts of mercy. They too may have a concept W that is linked to blame in the way envisaged. But are they being reason-responsive when they blame Theresa for her kindness?"

In elaboration of the worry, imagine a depraved moral code (the Klingon ethic) built around the preservation of personal honour and the elimination of impurity. We imagine that Sarpek the Klingon's opinions about which acts are immoral diverge utterly from ours, so that any time where we are both inclined to classify an act as "morally wrong" the convergence is purely accidental. Unlike my approach in 4.1 I am not inclined to make any further tweaks to the statement of stability to exclude Sarpek. Let us assume that he has false, repugnant views specifically about morality.

The Klingon-ethic worry is initially simply a challenge: does interpreting Sarpek's W as *moral wrongness* make him reason-responsive? The relevant premise in the argument (instantiated in the way relevant to Sarpek's case) is the following:

> (*) Any reason-responsive agent would be such that: the judgement that Theresa's kindness was wrong (and unexcused), makes them blame Theresa for her kindness.

When presenting the argument in section 2, I gave an instance like this for Sally, but then noted that to get the general stability conclusion we needed to generalize it to all agents and acts. The question here is whether that generalization was legitimate. Here are three rival descriptions of Sarpek's case:

(A) Sarpek (all things considered) ought not blame Theresa for her kindness---such feelings on his part would not be reason-responsive. Kindness is not something blameworthy! His moral judgements are not reason-responsive either. For example, perhaps (as Harman has argued) false moral judgements are *morally* wrong even if epistemically justified. [11]

[11] Harman (2011, sec 4) holds that we have a moral obligation to believe relevant moral truths. Applied to this case, the Harman position would be that Sarpek's wrongness-judgement was morally blameworthy (and I assume not overall reason-responsive) even if Sarpek were epistemically perfect (cf. p.462). I think Harman's case here is most plausible if restricted to cases of morally depraved beliefs, rather than ordinary moral error such as that of Sally in 4.1.

(B) Sarpek (all things considered) ought to make exactly the moral judgements he does. They are the only justified opinion to reach, given his anti-social upbringing and misleading experience, and pace Harman epistemic factors determine what one ought to believe. Further, though Theresa is not blame*worthy*---from an objective point of view, she ought not be blamed---Sarpek would not be responding to the reasons he has if he didn't blame her, given his course of experience.

(C) Combining elements of the above: Sarpek's moral judgements are reason-responsive (given his misleading course of experience) but is not justified in feeling blame towards Theresa.

Which description is correct is contested territory in first-order moral theory, and I won't try to adjudicate it here (it may vary on different ways of filling out Sarpek's case). I will trace the consequences of each for the present account. Descriptions (A) and (B) pose no threat to the wide-scope premise (*). If (B) is the correct, then morally depraved agents fit seamlessly into the story given in this paper. (A) is more threatening, since depraved agents will be *unreasonable* in both believing and blaming. Rival interpretations, on which W denotes something other than *wrongness*, could make him *epistemically* more reasonable, albeit at the cost of being unreasonable in linking *those* judgements to blame. On this description, depraved moral agents such as Sarpek might be like Suzy from 4.1. They would thus not pose a *new* threat to my argument, but they would make the Refinement to Stability introduced in section 4.1 more significant.

Description (C) is the only one on which (*) itself comes under pressure. The following triad is uncomfortable: (all things considered) it is reason-responsive to believe that p, reason-responsive to lack sentiment s, and yet reason-responsive to have a disposition to have feel s when one believes p. If we hold fixed the first two normative claims, there is pressure to give up the third claim (*). On reflection, though, I recommend we should not give it up even so. Instead, we should say that it is impossible, in Sarpek's situation, to be perfectly reason-responsive: he is condemned

either to believe inappropriately, blame inappropriately, or have an inappropriately run mental economy. If we retain (*) in this spirit, then the explanation of stability applies to Sarpek just as before. We make him more reasonable if we interpret him as having genuinely moral judgements connected to (*). Further, these moral judgments themselves would be reasonable, so unlike description (A), there is no countervailing pressure. He is, admittedly unreasonable in blaming Theresa for her kindness, but that doesn't look like a flaw that could be removed by judicious reinterpretion, so there's no *more reasonable* rival interpretation in the offing.

It is a theme of this essay that the metasemantic verdicts that substantive radical interpretation delivers turn crucially on first-order normative theses---often interesting and contestable ones. The case of Sarpek highlights some of these connections.

## CONCLUSION

I have offered a metaphysics of mental representation, substantive radical interpretation and shown how to derive referential stability (part 2). The rest of the paper has filled out the story, either by contextualizing it against related rival accounts or by testing it against objections. The take-aways are a refined understanding of the referential stability thesis, an explanation of why it obtains, and a sharp sense of the interaction of theses in the theory of content with theses in the theory of reasons.

The referential stability of wrongness is a problem for many extant theories of contents. The trouble for folklore Lewisianism covered in section 3 is a case in point. One can delay the day of reckoning by giving a partial account restricted to non-normative concepts, but ultimately one has to show that these partial pictures cohere with a general account of what it takes for concepts of all kinds to pick out what they do. I have made a case that substantive radical interpretation passes this test. Since I believe in referential stability, I see that as confirmation of this theory. I offer this as a contribution to the debate on the correct metaphysics for thought.

I also offer substantive radical interpretation as a contribution to the metaethical debate on the significance of referential stability. Notice: I have at no point taken a stand on the metaphysical status of *moral wrongness.* One could think of this as a property built up out of naturalistic properties, or (as Dunaway and McPherson suggest) a *sui generis*, elite property of serious ethical science; one could equally think of it as a non-natural property, or theorize it in expressivist terms (assuming that expressivism is not itself treated as a metasemantic view that competes with substantive radical interpretation as Ridge (2014) advocates). Nothing within the account depends on the metaphysical status of moral wrongness, but only on its role in first-order normative theory, as laid out in Part 2. Those assumptions are not truisms, and they require us to adopt particular stances in moral psychology and normative theory, but they are not metaphysical. So the account here constitutes a standing challenge to anyone who would try to argue that referential stability of wrongness can only be accounted for on their preferred metaphysics of morality.

**Bibliography**

Arpaly, Nomy (2000). "On Acting Rationally Against One's Better Judgement". *Ethics* 110, pp. 488–513

Boyd, Richard, 1979. “Metaphor and theory change”. In: Ortony, A. (ed), *Metaphor and Thought*. Cambridge: Cambridge University Press.

Boyd, Richard, 1982. “Scientific Realism and Naturalistic Epistemology”. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 1980 (2), pp. 613-662. East Lansing MI: The Philosophy of Science Association.

Boyd, Richard, 1988. “How to be a moral realist”. In *Essays on Moral Realism*. Ithaca: Cornell University Press, pp.181-228.

Dancy, Jonathan, 2000. “The Particularist’s Progress”. In: Hooker, B.W. & Little, M. (eds.), Moral Particularism (Oxford: Oxford University Press, 2000), pp. 130-156.

Dancy, Jonathan, 2004. *Ethics without Principles*. Oxford: Clarendon Press.

Darwall, Stephen, 2010 "But It Would Be Wrong", Social Philosophy & Policy 27.2.

Dowell, Janice (forthcoming). “The Metaethical Insignificance of Moral Twin Earth”. *Oxford Studies in Metaethics*.

Dreier, Jamie (1996). ‘‘Review of Smith: *The Moral Problem*,’’ *Mind* 105, 363– 67.

Dunaway, Billy & McPherson, Tristam (2016). “Reference Magnetism as a Solution to the Moral Twin Earth Problem” *Ergo* Volume 3, No. 25, 2016

Elstein, Daniel Y. & Hurka, Thomas, 2009. “From Thick to Thin: Two Moral Reduction Plans”. *Canadian Journal of Philosophy* 39(4), pp. 515-535.

Field, Hartry H., 1978. “Mental representation”. *Erkenntnis* 13(1), pp. 9-61. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, Oxford, 2001) pp. 30-67.

Gibbard, Allan, 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment.* Cambridge, MA: Harvard University Press.

Goodman, Nelson, 1954. *Fact, Fiction and Forecast*. University of London: Athlone Press.

Harman, Elizabeth, 2011. “Does Moral Ignorance Exculpate?” *Ratio* 24(4), pp. 443-468.

Hare, R.M., 1952. *The Language of Morals.* New York: Oxford University Press.

Heuer, Ulrike, 2010. “Wrongness and Reasons”. *Ethical Theory and Moral Practice* 13 (2), pp. 137-152.

Horgan, Terence & Timmons, Mark, 1992. “Troubles on Moral Twin Earth: Moral Queerness Revived. *Synthese* 92(2), pp. 221-260.

Lewis, David K. 1974. “Radical Interpretation”. *Synthese* 27(3), pp. 331-44. Reprinted in Lewis, *Philosophical Papers I* (Oxford University Press, 1983), pp. 108-18.

Lewis, David K., 1983. “New Work for a Theory of Universals’. *Australasian Journal of Philosophy* 61(4), pp. 343-377. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999), pp. 8-55.

Lewis, David K. 1984. “Putnam’s Paradox”. *Australasian Journal of Philosophy* 62(3), pp. 221-36. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999), pp. 56-77.

Lewis, David K., 1992. “Meaning Without Use: Reply to Hawthorne”. *Australasian Journal of Philosophy* 70(1), pp. 106-110. Reprinted in Lewis, *Papers on Ethics and Social Philosophy* (Cambridge University Press, 1999), pp. 145-151.

Mill, John Stuart, 1863. *Utilitarianism*, London: Parker, Son & Bourn, West Strand

Pautz, Adam, 2013. “Does Phenomenology Ground Mental Content?” In: Kriegel, Uriah (ed.), *Phenomenal Intentionality*. Oxford: Oxford University Press, pp. 194-234.

Putnam, Hilary, 1980. “Models and Reality”. *The Journal of Symbolic Logic* 45(3), pp. 464-482. Reprinted in: Benacerraf and Putnam (eds.), *Philosophy of Mathematics: Selected Readings*, Second Edition. (Cambridge University Press, Cambridge: 1983), pp. 421-446.

Putnam, Hilary, 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Railton 1986, “Moral Realism” The Philosophical Review, Vol. 95, No. 2 (Apr., 1986), pp. 163-207

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

Scanlon, 2007. "Wrongness and Reasons", in Oxford Studies Metaethics 2.

Schwarz, Wolfgang, 2014. “Against Magnetism“. *Australasian Journal of Philosophy* 92(1), pp. 17-36.

Schroeter, Laura & Schroeter, Francois, 2013. “Normative Realism: Coreference Without Convergence?” *Philosopher’s Imprint* 13(13).

Skorupski, John M., 2010. *The Domain of Reasons*. Oxford: Oxford University Press.

Smith, Michael, 1994. *The Moral Problem*. Oxford: Blackwell.

van Roojen, Mark, 2006. “Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument”. *Oxford Studies in Metaethics* 1, pp. 161-194.

Weatherson, Brian, 2013. “The Role of Naturalness in Lewis’s Theory of Meaning”. *Journal for the History of Analytic Philosophy* 1(10), pp. 1-19.

Wedgwood, Ralph, 2001. “Conceptual Role Semantics for Moral Terms”. *Philosophical Review* 110(1), pp. 1-30.

Williams, Bernard, 1981. “Persons, Character and Morality”. In: *Moral Luck* (Cambridge: Cambridge University Press), pp. 1-19.

Williams, J Robert G, 2016. “Representational Scepticism: the bubble puzzle” in *Philosophical Perspectives: Metaphysics*, supplement to *Nous*, 30:1 (2016) pp. 419–442