4

Reactive Concepts

Engineering the Concept CONCEPT

David Braddon-Mitchell

1. Introduction

This is a volume about the enterprises known variously as conceptual engineering[1] and conceptual ethics.[2] Exactly what is the nature of the relationship between these terms is still fluid; but a good way to rationalize it might be this. Conceptual engineering is the business of changing existing concepts and devising new ones. Conceptual ethics is the business of evaluating existing concepts and ways of talking, along with newly engineered ones, and making normative judgments as to whether they are fit for purpose. Jointly they promise reform through innovation and selection of how we think and talk in the pursuit of various ends—whether moral ends, practical ends, or alethic ends.

Thought and talk are the promised targets of these conceptual reforms. I'll set aside talk for now, though I'll come back to it. But 'thought' suggests—as does the word 'conceptual'—that at least in part concepts are a key part of the reformist's ontology. She will be engineering new ones, and recommending the adoption of some new ones, and the revision or sidelining of others. But what about the concept CONCEPT itself? Could there be ways of thinking about what concepts are that are different from standard understandings, but which have benefits that the conceptual reformer might take to be significant enough to merit recommending that we think and talk in terms of them?

This chapter recommends such a revision, or perhaps addition, to our stock of ideas about concepts. Classical and neo-classical accounts of concepts are thought to be connected in various ways to regular beliefs[3]—they enable us to have certain kinds of beliefs with certain kinds of content. The key idea in this chapter will be that there is a kind of mental concept which is connected to a mental state that is a little different from the usual conception of belief. Beliefs—straightforwardly in the functionalist tradition[4] but also in many others[5]—are individuated by what one might call

[1] Floridi (2011); Cappelen (2018). [2] Burgess and Plunkett 2013a.

[3] For a taxonomy of accounts, see chapter 1 of Margolis and Lawrence (1999).

[4] Braddon-Mitchell and Jackson 2007.

[5] Consider, for example, the covariational program that starts with Dretske (1981) or the teleonomic one which has perhaps its best expression in Shea (2013).

David Braddon-Mitchell, *Reactive Concepts: Engineering the Concept CONCEPT* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020).
DOI: 10.1093/oso/9780198801856.003.0004

their 'input' conditions. At its crudest these are just what distally causes them, but various traditions include what they co-vary with, what ideally causes them, and so on. At its most general, we might describe them as word-head connections. The connection to behaviour (and thus the head-world connection) is purely via their interaction with desires: distinct states which interact with beliefs to produce behaviour.

Let me give a motivating example which I will discuss later: an unspeakable word and associated concept which, in the case of racist speech in my culture (I'm an Anglo-Celtic Australian), represents an indigenous person. In order not to distract by mentioning the word (mention sometimes has spillover effects from use) I'll pretend that word is "Arthur" and the corresponding concept ARTHUR. What's racist about it? It's not just linguistic. It's that corresponding to that word there is a racist *concept*—or at least a mental entity in the concept family. And that mental entity does not merely represent skin colour or information about ancestry. Rather, it's one which when tokened itself *directly* causes negative affect of a range of types, which in turn both directly and indirectly motivates discrimination and other behaviours; none of this happens indirectly via discrete racist beliefs of the form 'Arthurs are lazy' and so forth. Recognizing that there are such mental entities in the concept family gives us the power to challenge them. If there is such a concept it's not one we should possess (although this chapter is of course asking you to possess the concept of such a concept—more on that distinction later as well). Perhaps, too, it will be explanatory: certain kinds of implicit bias may be explained by the fact that such a concept is possessed, even though no explicit negative beliefs are held, and the concept may be given a linguistic label the same as the anodyne representational concept (perhaps the tacitly prejudiced person possesses this special kind of concept ARTHUR while using the word 'Aboriginal'.

This chapter will recommend, then, that we consider a kind of concept which bears a relation like the one traditional accounts of concepts bear to beliefs, but instead bears it to states individuated not only by their causal inputs, but also by their *direct* causal outputs. I'll call these states **reactive representations**, **RR**s for short. They are partially representational states which are reactive inasmuch as they bypass interaction with distinct desires to directly motivate behaviour. In a later section I'll discuss how this relates to well known objections to 'besires': states that combine features of beliefs and of desires.

I will argue that there seems to be an important difference between two sorts of representational mechanisms: ones that really do act like mere beliefs where the representational state is discreet, and waits for a motivational state to come along and interact with it, and cases where the act of recognition—the act of representing—has itself *immediate* motivational force, as in the case of 'thick' moral terms,[6] certain sorts of phenomenal concepts and, I think, cases of bias and hate. When the motivating force is immediate, the motivation itself becomes part of how we taxonomize the

[6] It might seem that by explaining thick moral terms and slurs in terms of concepts I'm coming down firmly on the side of a debate about whether these are pragmatic or semantic phenomena (see Väyrynen 2013) for a defence of the pragmatic account). Certainly that's one way to taxonomize the view I will defend. But another would be to somewhat problematize the pragmatic/semantic distinction.

world: being apt for motivating in that way is part of what makes us see something as falling under the concept. I'll call a concept that is connected to an RR in the way an ordinary concept is connected to an ordinary belief a **Reactive Concept,** or **R-concept**. Whether there are such things, as I understand them, is an empirical matter which my research team is investigating. Here I argue only that the hypothesis that there are does a lot of explanatory work, and most importantly for current purposes, makes sense of the focus on the reform of concepts on normative grounds that has been called conceptual ethics (Burgess and Plunkett 2013a,b).

2. The R-concept Hypothesis

In this section I will first elaborate a little more on the hypothesis that there are R-concepts and then consider how they compare to ordinary concepts, ordinary beliefs, and desires. I then go on to consider whether the idea can be easily subsumed by global belief-desire psychology.

2.1. Concepts, R-concepts, Beliefs, and Desires

Exactly what the relationship is between beliefs and desires and representational content is complicated. But for our purposes here let it be assumed that concepts are part of what makes it possible to possess certain beliefs or desires, or what makes it possible to token a representation with a certain propositional content.[7] Insofar as you possess the concept 'triangle' you can token states that represent triangles. So perhaps to possess a concept is to possess a kind of ability. There are complications that don't matter for now: there is a tradition according to which a concept is what allows one to explicitly and consciously represent a certain state, others on which they are required to represent something *de dicto* rather than merely *de re* (on such views mental states might have non-conceptual content[8]). But the simple formulation will do for our purposes.

Concepts traditionally, then, fall firmly on the representational side of the mental family. If I possess the concept 'possum' then that is *part* of what it takes to be able to form at least *de dicto* beliefs about possums[9]. Rational people require, in addition, evidence of various sorts to form beliefs. In the case of beliefs, possession of the concept together with evidential or causal connections makes, in normal subjects, possession of the belief in some cases almost automatic. If I have the concept POSSUM and I am presented in experience with the features of possums that are encoded in POSSUM then at very least I will form the belief that it seems that there is a possum here. If I take myself to be in a position to trust my perception, then I'll form the belief that there is a possum nearby.

The idea here is firmly attached to standard belief-desire psychology. Beliefs are states which when formed have no power to motivate by themselves. Rather they sit in the head, waiting to be paired up with a state which will interlock with them. Their

[7] This is close to the rough characterization in Burgess and Plunkett (2013a) except that it's a necessary condition, rather than a sufficient one for thought formation.

[8] See Bermúdez (2007) and Roskies (2008) for some up to date discussion of this idea.

[9] And desires, but the point which follows about automatic production may not apply in this case.

semantic values, while they depend on the overall cause roles they play, perhaps together with desires, are invisible to the desire mechanism.[10] Once the belief that, for example, a possum is present is formed, a mental entity hangs around with a local syntactic shape that can be recognized by desire-like objects. If the desire to be near a possum encounters a belief that a possum is near, then (*ceteris paribus*) the agent doesn't move. If it is encountered by a desire for possum flesh, quite other behaviour is caused.

It is not part of the individuation condition of a belief that it causes any particular behaviour; the behaviour depends entirely on the agent's desire profile. And a standard concept is, inter alia, something that is required to form beliefs.

But notice that it's a contingent fact that the *token* physical state which is sitting around waiting for desires doesn't itself directly cause behaviour. In principle there could be such states; states formed by certain input conditions, which indeed might sit around and wait for desires, but in the meantime directly shortcut the belief-desire system to produce behaviours. There could be a state, co-varying and generally having whatever your favourite foundational semantics says makes it about possums, which directly motivates running away from possums. Let's call this state 'UGH-POSSUM'. The state itself, when formed, has the power to initiate movement away from the possum. Insofar as it has belief like features, it can also combine with other desires about possums (if you independently desire to sing a silly song about possums whenever near one, so you will sing as you leave). But it doesn't need any such extra desire to do *certain* things. This token state realizes[11] a belief that there is a possum nearby. But it does not *only* realize a belief that there is a possum nearby: it has extra powers that not every belief that there is a possum nearby would have. A kind to which this token state belongs is not individuated by the purely representational semantics, it's individuated as well by its output powers: it's powers to directly produce behaviour. Such a state is what I'm calling a reactive representation. An immediate concern that you might have is that this sounds an awful lot like a besire, and there are well known objections to the very possibility or coherence of such states. But nothing I say here depends on showing those objections don't work, as in section 4.2 I argue that in fact RRs are not besires.

The extension of the idea of a concept I'm recommending here bears the same kind of relationship to RRs as the traditional idea of a concept does to beliefs. The RR 'UGHPOSSUM' is a state which is belief like inasmuch as it responds to inputs; is caused by possums; and tracks perceptual evidence of possums. But rather than waiting for desires to come along and interact with it, it directly induces aversion of the detected object. The R-concept UGHPOSSUM is a kind of ability. It's the ability to form these states. Possession of the R-concept UGHPOSSUM is required to have the ability to form the RR 'UGHPOSSUM'. And just as in the case of regular concepts, where there were conditions in which possession of the concept makes forming beliefs automatic, there will be conditions in which possessing a certain R-concept will make tokening the relevant RR automatic. If someone who possesses a certain

[10] A point vigorously stressed in Fodor (1987) and countless subsequent publications.

[11] The reader wondering why I have said 'realizes' rather than 'is' will have her questions answered in section 4.2.

R-concept sees presentations as of the representational features encoded in the R-concept UGHPOSSUM then they will token the RR UGHPOSSUM, and this will in turn produce the downstream causal effect of fleeing the marsupial infested neighbourhood.

2.2. Comparison with Standard Accounts of Concepts

So in using R-concepts to taxonomize the mind we will draw distinctions differently. Two concepts that have the same input conditions will count as the same ordinary intensional concept. But if two states which have the same input conditions, but one gives the capacity to form regular beliefs, and the other gives the capacity to produce RRs which bypass the standard desire mechanisms to motivate behaviour, they count as different along the R-concept dimension. Is there any motivation to use the term 'concept' for R-concepts at all, rather than engineering an entirely different term for this sort of mental entity?

There are plenty of *prima facie* reasons for using the word 'concept' here: clarity, ease of explanation, and so on. But another is that we already need to vary the standard intensional individuation of concepts to explain mental phenomena which are amongst those whose explanation it is the job of a theory of concepts to provide.

Ordinary concepts are often thought to be fully individuated by their input conditions—or at least what it takes to fall under them is. But something other than the 'ordinary' idea of individuation of concepts by the features of the things which makes them fall under that concept may already be needed to account for merely hyperintensionally distinct concepts.[12] Some such hyperintensional distinctions might, for example, be marked by *method* of detection, rather than features required for the thing to fall under the concept. It takes the same features in the world for something to count as *falling under* TRIANGLE and TRILATERAL, but one can imagine someone disposed to try to identify these figures by counting angles insofar as she is deploying the TRIANGLE concept, and angles insofar as she is deploying TRILATERAL.

So are TRIANGLE and TRILATERAL the same concept or are they different concepts? Surely there is no need to make that call: they are intensionally the same but hyperintensionally distinct. If one wanted terminology, one could say they are same concept but different H-concept (friends of the hyperintensional often like to think that concepts are individuated hyperintensionally, and they use another world for intensional concepts, but that's a fight about words).

This seems to be the right thing to say in this case, and no-one I know of has suggested that insofar as concepts are hyperintensionally distinct they are not conceptually distinct in some sense (if anything, as above, the other move is more often heard—that concepts are hyperintensional through and through, and that intensional distinctions are coarser grained than the usual conceptual ones). And if it is the right thing to say in this case, then it seems to apply to the case of R-concepts as well. For here also we have intensionally (and maybe hyperintensionally) distinct

[12] For a sketch of an account of this that does not invoke metaphysically robust impossibilities, see Braddon-Mitchell (2009).

states, which because they function causally in slightly different ways, get to count as different in a different dimension from the intensional or hyperintensional.

2.3. R-concepts versus Belief Desire Psychology with Irrationality or a Divided Mind

So if the distinction between R-concepts and standard concepts is a good one, the easy job of saying that they deserve to be thought of as a kind of concept is done. But our comparison with standard accounts of the relationship between beliefs and concepts is not yet done. For there is a danger that R-concepts and RRs might be redundant given the way belief-desire psychology is often understood.

So far I have introduced the idea of RRs in terms of the different behaviour of token concepts. The thought is that there are tokens of the required types in the head, and there is a difference in behaviour between RRs and standard beliefs. Standard beliefs stand as if at a dance, looking for a desire partner with whom to generate behaviours. RRs do the work themselves, as it were, without a partner.

But on a global conception of belief-desire psychology this doesn't make sense. On this conception, thinkers don't have individual beliefs and desires. Rather they have global input output profiles, complete belief-desire states (Braddon-Mitchell and Jackson 2003: chapter 3). Individual beliefs (or desires) only make sense as abstractions that describe the minimal differences between global states.

On this conception having the R-concept UGHPOSSUM is indistinguishable from having an overall belief-desire profile which includes the belief that there are possums around, and a desire to vacate the possum rich environment. For you are globally set up such that when you encounter possums you flee them.

For many purposes that's the right conception of belief and desire, in my view. But I'm less sure than I once was that it's *always* the most useful conception. In particular, it doesn't play well with typical actual behavioural profiles in that it can't make sense of irrationality.

Suppose, for example, that you are disposed to apparently randomly flee possums or not. It turns out that what is going on is that you approve of possums; you have a standing desire to observe them. But nonetheless you possess the UGHPOSSUM R-concept, so your categorization process produces an immediate impulse to flee. At times when you are reflective, you overcome this because of your desire to observe and be in the company of possums. At other times you just react unreflectively and flee the neighbourhood. Or consider someone who sincerely claims that certain behaviours are likely to cause eternal torture, who sincerely believes that infinite torture massively outweighs fleeting pleasure, yet performs behaviours which would, if this were true, bring about the infinite pain and only modest pleasure—and does all this with only modest regret. The global story here has unilluminating fixes:[13] you have context dependent desires (you desire-when-reflective to be with possums, you desire-when-unreflective to avoid them) or the divided mind hypothesis (Lewis 1982; Stalnaker 1984; Davidson 2004; Egan 2008). It's unsurprising that these are a little unilluminating about the exact behaviour of mental states in less that perfectly

[13] For a nice account of the ways in which these are unilluminating, see Norby (2014).

rational beings like us: for all of these fixes are not designed to taxonomize the mind in terms of token states, but rather to preserve *global* accounts of the way the mind is, removing contradiction either by creating two separate consistent accounts (the fragmentation story), or one in which our desires and perhaps beliefs are so massively context dependent they don't play the kinds of roles in our ordinary discourse that they were intended to (e.g., a desire-when-reflective as against a desire-when-unreflective, or a belief-when-talking-about-the-act as against a belief-when-acting).

When we aren't talking about an abstract global theory of what it is to be a rational being (for which global belief-desire theory seems eminently suited) but rather interested in *token* explanations, the hypothesis that sometimes a single token state both motivates and represents, and at other times the work is divided between distinct token states, is both illuminating and empirically tractable. It seems to explain a lot if true, and it shouldn't be beyond our powers to find out if it is true.

3. General Explanatory Benefits of R-concepts Including Ones for Conceptual Ethics

In this section I argue that, if the R-concept hypothesis is correct, it would not only provide direct benefits in explaining the importance of conceptual ethics but would also provide general explanatory benefits.

3.1. Three Unified Benefits: Hate Speech, Crypto-evaluative Terms, Phenomenal Concepts

Here's another benefit of the idea of the R-concept. It provides a unity of explanation over a range of three apparently very different areas. I'll briefly survey the advantages that this story might have in each of the areas, suggesting that the unity over the areas it brings is a theoretical advantage, and that in each case understanding that there are R-concepts involved makes better sense of the projects of conceptual engineering and conceptual ethics in these domains.

3.1.1. HATE SPEECH

I've already given the example of hate speech: here the idea is that to hate speech there might correspond hate R-concepts; which give you token states which both represent the external world a certain way, and cause you to react in a certain way. That the reaction is so immediate also gives the impression that being apt for the reaction is part of what it takes to fall under the R-concept. So if we wanted a theory of content for R-concepts, then this would likely come with the territory. The R-concept *ARTHUR causes you to token states which represent indigenous Australians, to respond in a certain way and also to form the view of them being represented as rightly producing that response*. You do not truly fall under that concept, the racist is likely to think, unless you are deserving of the response. This makes richer sense of the recommendation that a practitioner of conceptual ethics might be doing when she recommends against using the word "Arthur". If it was just an issue of the *word* then the focus seems to be on two possible worries. One is not using the word because it gives offence. This is a weak justification for two reasons. One is that there is a

liberal conception of offence as a very weak kind of harm, which hardly justifies restriction of this kind of speech unless we extend the harm into a kind of verbal violence. Perhaps a reasonable move, but not an uncontroversial one. The other is that it gives *evidence* of racist attitudes on the part of the speaker. But this would be superficial evidence: if the underlying *concept* is just the same as the representational concept "Aboriginal" the only racism is being careless with possible offence. But the intuition many have is that it's not an unqualified benefit to have racists hide their colours by careful choice of words. So there's something more going on than just choice of words. On the R-concept story, though, it is indeed evidence: but one which shows that the user of the word may possess a certain R-concept, something which other things being equal disposes them to prejudicial action. And making a recommendation against such a concept, even to someone in the grip of it, might sometimes help. Seeing how hard it is not to talk in those terms, she can become aware of it and, if in respect of her considered desires they are reasonable, she may try to remove the concept from her mental repertoire. I discuss in section 3.3 the prospects for finding a connection between use of words and R-concepts.

Perhaps here a few words are appropriate to discuss the extent to which accepting my view requires you to take sides on existing views about pejorative language. You might think that my view, tied as it is to R-concepts, falls firmly on the side of analyses which place the emphasis on the content of the expressions (see, e.g., Hom 2008). But that of course depends on what you mean by 'content'. I talk of the 'content' of R-concepts, and call R-concepts a kind of concept because I think that there are similarities in the way these notions do explanatory work to purely descriptive concepts and contents. But of course it's an *extension* of both ideas. And as such it's ultimately terminological (though not therefore unimportant) whether the extensions deserve the original terms. Another way of using the ideas I propose would be to accept that the mental entities I call "R-concepts" exist, and think of them as the underlying mechanisms that explain much of the pragmatics of thought and talk. The view is inconsistent, though, with any account according to which there is nothing mental that distinguishes what's going on in these cases from cases where there is no pejorative or hateful language. A very flat-footed version of Anderson and Lepore's view in their (2013) according to which the fact that the words are taboo is the *only* explanatory factor would be inconsistent with it, and the explanation would be entirely outside the head of the user. Finally, much of the literature concentrates on things that have speech does. There is a tendency in philosophy, as elsewhere, to promote single factor explanations, so that the promoter—should she win—gets to have a monopoly on being right. To the extent that linguistic behaviour has many important effects, one might think that this should flourish. So, for example, Camp's view (2013) that pejorative language serves to signal alignment with a certain perspective is something, which if just baldly stated like that is likely often true, and certainly consistent with what I say. Even stronger versions of it—such as the view that the distal explanation of its prevalence is to do with such signalling—are more empirically risky, but also consistent. What isn't consistent would be the view that none of the proximal psychology is as I describe.

3.1.2. PHENOMENAL CONCEPTS

At first blush, it might seem as though phenomenal concepts have very little to do with the topics we have been discussing. But I think that R-concepts explain important features of them as well. Consider a discussion which went on over many years between me and my friend the sadly late philosopher Jonathan McKeown-Green. Jonathan was blind from birth, and as a philosopher interested in perception, he was a natural person to talk to about colour concepts. At some level he thought he couldn't possibly have the same concept RED that I had. But intensionally he possessed a concept (let's call it Red-J) that was identical to mine: let's say it tracked a certain reflectance profile. And yet, he and I thought, it wasn't the same concept.

When there is similarity at the intensional level, but difference of concept, you might diagnose hyperintensionality. But that's not what's going on here. If one of our concepts was hyperintensional, it would categorize in a more fine grained way. But there's nothing more fine grained here in either of our concepts.

Perhaps it's response dependence that's at issue (Pettit 1991). But moving to a typical response dependent conception doesn't help either. If RED is to be characterized as "apt in normal conditions to produce response R in species-typical individuals" then again we have an ordinary intensional concept, and something that Jonathan and I could agree on, while not dispelling the sense that we don't possess the same concept—that somehow, while we agree on this, he didn't understand what sighted people *mean* by a term like 'red'.

But notice that there is an R-concept and RR in the neighbourhood. Recall that an RR is something which is responsive to information about the world, but part of its individuation is via not just those inputs, but its outputs—the things it causes. So far the outputs we have considered are ones of reactive attitudes and behaviours. But reactive phenomenologies are just yet more causes.[14] To possess the R-concept RED is to possess a capacity (and perhaps a disposition) to token the RR 'red'. The RR 'red' is tokened when the agent receives certain information about the world[15] but also has the effect of producing the phenomenology.[16]

This, then, is something Jonathan and I do not share. He does not possess the R-concept RED as he does not possess the ability to token the RR 'red', even though he does share the intensional concept RED. The R-concept RED has as a feature an effect on what it takes to be RED in this sense—that is, had produced this effect in me. So while there is a sense (derived from the intensional concept RED) in which we can

[14] By phenomenologies here I don't mean anything substantive. If you are a dualist who believes in rich qualia, let them be the causal product of tokening certain RRs. If you are at the other extreme, and think it's all about dispositions to make verbal claims about experience, let those dispositions be possessed by the RR. And for every positions in between the same can be said.

[15] If this were a chapter about phenomenal concepts, then some discussion of whether the input channels are relevant here would be appropriate—maybe it makes a difference if it is tokened by the perceptual system.

[16] There are interesting issues here about how integrated an RR has to be. Perhaps empirically we may find the cause of phenomenal component in the perceptual system, and the cognitive component elsewhere. And yet somehow there is a "binding" of the two. The bound pair might be thought to be a cause of the phenomenology, but possibly not the cognitive component alone.

both judge that some flower is red (me by looking, he by consulting a colour meter) there is another in which he cannot judge that it's red in the same way as me.

3.1.3. CRYPTO-EVALUATIVE CONCEPTS

Finally a certain kind of puzzle in giving an account of certain concepts is explicable if there are R-concepts in the vicinity. These are concepts I call 'crypto-evaluative'. They might include HAPPINESS, FREEDOM, DEMOCRACY, and many others. I call them crypto-evaluative because on the surface they seem amenable to analysis in terms of what features of the world are required for them. We might, for example, give a story about what decision making processes count as democratic. But they are all vulnerable to normative objections that expose the hidden evaluate features. Suppose an agent considers an excellent account of what democratic decision making amounts to (and thinks herself a democrat) but judges that that's not a good way to organize society. Likely she will make the 'true democracy' move: she will be less inclined to judge that that procedure really is democratic.

Consider the debate about the concept of happiness. Various descriptive accounts of happiness—ones which feature objective lists, or which feature psychological states—have been criticized on the grounds that these things could in principle turn out to be undesirable. (Nozick 1989; Nussbaum 2008). The obvious alternative is explicitly normative accounts: happiness is the greatest prudential good, or happiness is flourishing of some kind. These in turn are criticized as leaving out a substantial account of what happiness is (Haybron 2008; Feldman 2010). (Maybe the concept is one which is committed to mental ease, or some other psychological state, as part of what it takes to be happy—no-one is happy, such a response goes, however much she has the objective list if she is miserable.)

If there is an R-concept HAPPINESS to be had (perhaps in addition to a standard one) then what that R-concept amounts to is the ability to form the RR 'happiness'. And, perhaps, the RR 'happiness' is tokened when there is a state which is caused by (or otherwise represents) some list of features—perhaps in this case internal psychological ones, perhaps also elements of an objective list—but which also causes motivation towards these features. So no-one would judge that she is happy (insofar as the R-concept is playing a role in the judgement) unless her representation of these states motivated her towards it. Thus we get an explanation of how someone can both be prone to think that there is substantial empirical content to the concept of happiness, and yet be drawn to the idea that some normative features are essential to it. Crytpo-evaluative concepts are just R-concepts.

3.1.4. TACIT BIAS

I'll add one very brief speculation. A lot has been written about implicit bias; bias where agents seem to have both explicit beliefs and explicit desires which don't appear to have any bias towards a group of people. And yet, the agent behaves in a biased way: perhaps being more inclined to vote against someone from a minority group in a job selection, for example. One hypothesis worth testing would be that R-concepts are involved here. If the biased motivation is part of the R-concept—a direct causal product of the representation—and not a result of stand alone desires metaphorically looking for beliefs to pair with, then it's more likely that such a

pattern might be something the agent herself could miss when interrogating her discreet beliefs and desires. There may be no stand alone desire not to see members of that group employed, and no stand alone belief that they shouldn't be, but the R-concept makes it likely that, confronted by members of a group, an RR will be formed which directly causes the biased behaviour.

3.2. Conceptual Difference Where There Is No Disagreement About Fact

So far I have done some setup about the idea of the R-concept, and discussed some of benefits of having the notion to hand. In this section I'll talk about another case that I think makes R-concepts worth exploring for conceptual engineers.

Sometimes there seem to be disagreements which matter a lot, but which don't seem to depend on any disagreement of fact.

Consider the concept of 'survival'[17] It's very contentious what exactly survival amounts to. I'm tempted myself to think that it's a difficult notion, which admits of degree (Braddon-Mitchell and Miller MS). But for current purposes we need only consider two alternatives; a physical continuity theory and a psychological continuity theory. The simple physical continuity theory is just one that says you survive over some interval iff you are related to the entity at the end of the interval by the ancestral of the physical similarly relation. That relation in turn holds when there is object causally connected to you which overlaps with your current self physically to a very high degree. The psychological continuity theory is much the same, substituting psychological similarity for physical similarity. It's usually assumed that our account of psychological similarity is one which allows psychological similarity across discontinuous physical change—as with material destruction followed by reconstruction, or downloading.

It's notorious that people disagree vigorously about the plausibility of these two theories, which in part is what makes them good things to discuss in introductory courses.

I've tracked the opinions of undergraduate students on these topics for a number of years, and found that roughly speaking 40% of classes of 400 students or so are attracted to psychological continuity theories, and 45% or so are attracted to physical continuity theories, at least insofar as they take this difference to explain their different reaction to teletransporter cases.

The well known teletransporter cases are ones where there is bodily destruction but recording of the psychological information, followed by bodily recreation at a distant location, down to the relevant degree of detail required for the psychological information to be preserved.

Those who are attracted to the psychological continuity theory say that teletransporters, should they come to exist, will provide painless easy travel, and will be happy to use them if the price is right. Physical continuity theorists are genuinely baffled by this, and tend to scream at their fellow students of the other persuasion "you don't get it: you are being paid money to be killed".

[17] The *locus classicus* is of course Parfit (1984).

Now the question that I'm hoping R-concepts can help with is this: what is the disagreement about?

This question arises because it is possible to divide up the sample of students in ways that factor out obvious sources of disagreement.

Here are some obvious sources:

(1) Perhaps some students believe in souls, and really hold a soul continuity theory. But they think that souls are directly connected to bodies, at least insofar as boldly destruction severs the soul from the earth.
(2) Perhaps some students are dualists, and they think that no matter what I try to tell them about how a recreated brain will support consciousness, they don't accept this, and thus think they'll no longer exist after teletransportation, because there'll be no experience.
(3) Perhaps students think that there are substantive metaphysical facts about what survival is grounded in. Survival is worth having, they think, and it is a natural feature of the world (other things people might call survival are mere gerrymandered things of no interest) and the job of metaphysics is to find what grounds survival: so these two possible candidates, the two continuity theories, are alternative accounts of what the grounds of survival are. The disagreement is about the metaphysical fact about the nature of the grounds of survival.

So in order to eliminate these possibilities as sources of disagreement, I excluded from my sample people who believe in souls, people who are dualists, and people who believe in grounding. Each of these groups have an answer to what the disagreement of fact might still obtain between people who agree on the fundamental physical facts. The people who believe in souls think that while we might agree about the physical body and its constituents, there is something else—the soul—which might or might not be present, and we could be tacitly disagreeing about that (your body will survive, and your brain, but not your soul). Dualists, very similarly, might take the view that the dualistic component of the mental may survive some (identity preserving) physical changes, but not others. And finally, these undergraduates were taught about metaphysical grounding: the version in which grounded entities are ontologically distinct yet grounded in underlying physical nature. On many such views it's a substantial metaphysical fact which fundamental states ground the non-fundamental, and if persons are non-fundamental, there can be a substantial disagreement as to which physical states ground them.

On eliminating these people whose views take them to think there is a substantial disagreement there was only a slight change. Of the remaining group (luckily for current purposes my university is in a culture with few soul believers amongst undergraduates) the split is now about even. And now the puzzle is more serious. All of them agree, for the sake of the example, on all the base metaphysical facts. They agree about the nature of physical continuity. They agree about the nature of psychological continuity. They agree about background physics and science more generally insofar as it is relevant. They agree that there are no substantial grounding relations which connect more abstract things like 'grounding' with the basic facts in the universe. So what is left?

The remaining obvious possibility is something like merely verbal or lexical disagreement. They disagree about what 'survival' means. But that seems too thin a thing to justify the reactions. The physical continuity theorists still hold that it's mad and bad to teletransport, because it doesn't preserve survival. If it were a mere lexical disagreement, then you'd expect that it could be solved by dictionary makers. And yet no physical continuity theorist thinks she would be dissuaded by the discovery that the *OED* says that the psychological continuity theory is the right definition of the English word 'survival'. Indeed, not all physical continuity theorists think that their own theory *does* give the right account of the English word, or that it matters that it does.

Parallel with a lexical story is what one might call a "mere" conceptual story. The disagreement is about which concept to possess. But the problem with this is that *both parties already possess both concepts*. Neither party thinks that one shouldn't possess the other concept. Indeed the physical continuity theorist couldn't think about what she is right about were it not for the fact that she possessed the concept 'the thing that is preserved in psychological continuity, and that PC theorists call "survival"'.

So what is the disagreement about, if it's not fundamental metaphysics, grounding, semantics, or concepts?

Well of course the striking difference is a difference that looks like it's not one of belief, but of motivation. The psychological continuity theorist is very much motivated to preserve psychological continuity at almost any cost. The physical continuity theorist is similarly motivated with respect to physical continuity.

So what's going on here might be the possession of different R-concepts. The teletransporters (those who believe they survive teletransportation) have an R-concept which takes facts about psychological similarity as input, and *directly* motivates behaviour seeking psychological similarity at all costs. The others have an R-concept which takes facts about physical similarity as input, and similarly motivates behaviour seeking *physical* similarity at all costs. The first category is motivated to provide benefits to their psychological successors; the second will only be so motivated if they are physical continuers.

How does the R-concepts strategy help? If it were merely a matter of each side using different ordinary concepts, it would be still mysterious. For as I will explain in section 4.1, successful explication of an ordinary concept to an agent entails it's possession by that agent, so both sides would possess both concepts. But as we will see, R-concepts are different: you can explain the R-concept ARTHUR successfully to someone without that person thereby possessing it (even if she possesses a kind of meta-concept, as we will see). So it's quite possible for the groups to possess different R-concepts, for this to explain the difference, and for that fact not to change even when we explain the idea of the R-concepts to both sides.

3.3. R-concepts and Words: How to Engineer R-concepts

A principal part of the justification for a theory of R-concepts is not just that it's a better way to taxonomize parts of our cognitive economy, but that it helps in the task of conceptual revision under the guidance of conceptual ethics.

The thought so far is that it gives us as conceptual ethicists a story of why certain R-concepts are bad ones, and as engineers, a motivation to engineer new ones and try to persuade people to adopt them.

But short of as yet practically impossible, and likely highly immoral, direct neural intervention with our conceptual structures, the principal handle we have on our conceptual life—at least our more complicated concepts—is via words. So it may pay to consider the relationship between concepts and words, and R-concepts and words.

The use of words amongst those who understand them tends to token the concept required for their understanding. If I say to you "be alert for triangles" you'll be more likely to look at your environment in a way which is alert to their presence, and thus form beliefs that there are (or aren't) triangles in the area.

The interesting question for our current concerns is how the word-concept pairing works in the case of R-concepts. I suggested earlier that R-concepts may play an explanatory role in the case of thick moral concepts. The question remains whether the word-concept pairing is sufficiently robust that influencing verbal behaviour can be part of an intervention strategy to minimize the harmful effects of harmful R-concepts.

If it were, then we should expect something like this. The word "Arthur" is associated with the R-concept ARTHUR. The phrase "Indigenous person" is associated with the R-concept INDIGENOUSPERSON. If you ask someone (who is neither a committed racist nor alert to conceptual and linguistic contributions to discrimination and oppression) to "look out for Arthurs" then, on sighting an indigenous person, you might expect her to be more likely to form an RR with the content ARTHUR which in turn produces automatic fear responses and general negative affect. If on the other hand she is primed differently, and asked to look out for indigenous people, then on sighting such a person, you might expect her to be more likely to token a representation only of the regular concept INDIGENOUS PERSON, and thus have more neutral responses.

Of course all of this is tricky, since word-concept pairings are not entirely robust. Perhaps part of what it is to be a committed racist, aside from explicitly held racist beliefs, is to have the racist concept ARTHUR so firmly entrenched that all words for indigenous people will token representations with the content ARTHUR, either directly or via a very short chain of inference.

This, of course, is a second empirical matter concerning R-concepts which needs investigating: not only the crucial one of whether we actually possess them, but also the way in which the tokening of RRs under their influence is, or is not, mediated by related words.

This is perhaps a good point to add one further speculation. A lot has been written about implicit bias; bias where agents seem to have both explicit beliefs and explicit desires which don't appear to have any bias towards a group of people. And yet, the agent behaves in a biased way: perhaps being more inclined to vote against someone from a minority group in a job selection, for example. One hypothesis worth testing would be that R-concepts are involved here. If the biased motivation is part of the R-concept—a direct causal product of the representation—and not a result of stand alone desires metaphorically looking for beliefs to pair with, then it's more likely that

such a pattern might be something the agent herself could miss when interrogating her discreet beliefs and desires. There may be no stand alone desire not to see members of that group employed, and no stand alone belief that they shouldn't be, but the R-concept makes it likely that, confronted by members of a group, an RR will be formed which directly causes the biased behaviour.

4. Two Objections to the Theory of R-concepts

4.1. Failing to Possess the Concept and Failing to Understand the Concept

Here is a *prima facie* puzzle. With regular, purely representational concepts, understanding a concept is roughly equivalent to possessing a concept. Once I understand that the concept of TRIANGLE is the concept of a plane figure with three sides, it follows that I automatically possess the concept. Understanding what that concept is gives me that concept. So by understanding what the concept TRIANGLE is, I will end up with any of the powers that are associated with the concept: the ability to form thoughts about triangles, for example. By understanding the concept MOTOR CAR I come to possess the concept, and form the ability to form thoughts about them. But by understanding the R-concept ARTHUR we do not, we hope, thereby possess it, and do not have the ability to form the relevant RR. This fact might make one think that an R-concept is not really behaving like a concept.

It is helpful here to distinguish between a concept, and a meta-concept: the concept of that concept. In the case of ordinary concepts they are distinct, but for most purposes that distinctness doesn't matter. The meta-concept of triangle is, for example, the concept of some kind of mental entity or power that enables one to form thoughts about three sided plane figures. That's not the same as the first order concept TRIANGLE. But acquiring the meta-concept is sufficient for acquiring the concept, so it rarely matters. Perhaps it matters in animal cognition or developmental psychology, where plausibly some animals, and humans at some stages of development, can possess concepts but not meta-concepts.

But the distinction matters here in realm of R-concepts. A possible complaint about the idea of an R-concept is that the idea that we might be able to resist such concepts is absurd—conceptual engineering will never be able to bring it about that we do not have these concepts, especially not if we go about explaining them. For typically, to explain a concept successfully is to bring it about that the person to whom you have explained it possesses the concept. If I do a really good job of explaining an alien concept to someone, perhaps the concept associated with the Danish word "Hygge"—a concept of cosiness and domestic wintry comfort which is at the time of writing fashionably held to be a key contributor to wellbeing—then I will have succeeded insofar as she now possesses the concept. So if I somehow explain the R-concept ARTHUR or the R-concept perhaps associated with the word for which "The N word" is a euphemism—then it might seem that I will bring about its possession. Then, as an objection, this can cut one of two ways. Either this is true, and so there's no benefit to talking about R-concepts, or else it's an objection to the possibility of there being any such thing as an R-concept.

The latter horn of the dilemma goes this way:

(1) (Premise) if R-concepts are concepts then explanation will result in possession.
(2) (Premise) Possession of R-concepts has motivational force and casual power.
(3) (Premise) Mere explanation cannot bring about by itself motivational force.
(4) (from (2) and (3)) Explanation will not bring about possession of R-concepts.
(5) (Conclusion—from (4) and (1)) R-concepts are not concepts.

The conclusion comes from applying *modus pollens* to the conditional in the first premise. But of course this is not a compulsory move. Instead we should consider how generally true the first premise is.

We saw that the first premise is indeed true of purely representational concepts. But it is not true of R-concepts. But this is because the relationship between first order concepts and meta-concepts is different in the case of R-concepts.

Mere explanation about the nature of an R-concept will create a meta-concept. But it will be a purely representational meta-concept. Not an R-concept. And purely representation meta-concepts do not have the automatic power to create the first order R-concepts they represent, in the way that they do have the power to create first order representational concepts.

Consider our racist example again. Giving lots of information about the R-concept ARTHUR will tell someone that there is a mental state which responds to evidence of Aboriginality, which is able to interact with standing beliefs and desires, but that without any such interaction motivates fear, distrust, and dislike. This information creates in a comprehending recipient, a concept, for sure. It creates a concept (a representational concept) of the R-concept ARTHUR. So it's a representational meta-concept. But having the concept of such a concept does not allow you to form thoughts of the kind "there's an Arthur". Because thoughts of that kind require you to have an RR, usually brought about by having the R-concept ARTHUR. And indeed nothing purely informational will bring that about. Of course there are meta R-concepts too: these might be concepts of R-concepts which come with built in motivational force for the acquisition of the first order R-concept. Perhaps these sorts of meta R-concepts may have some power to bring about possession of the R-concept, but these meta R-concepts cannot be acquired by purely informational means either.

So premise (1) should be rejected. It's not true of all concepts that explanation of them results in possession of them. Of course this does show that there is a difference between purely representational concepts of which this may be true, and R-concepts for which it is false. But that's unsurprising. R-concepts are different from regular concepts, and this difference exactly tracks the key, motivational, difference.

4.2. Reactive Representations and Besires

There is a substantial body of literature which denies the possibility of so-called **besires**: states which are both beliefs and desires (e.g., Lewis 1988, 1996; Price 1989; Smith 1994; Zangwell 2008). R-concepts are defined in terms of reactive representations: and RRs at first blush look like they play both the belief role and the desire role.

This would make them a kind of besire. There can be no R-concepts without RRs, so if there are no besires, there are no R-concepts.

Let's lay this out:

(1) (Premise) R-concepts are defined in terms of RRs.
(2) (From 1) If there are no RRs there are no R-concepts.
(3) (Premise) RRs are a kind of besire.
(4) (Premise from the Besires literature) Besires are impossible.
(5) (3 and 4) There are no RRs.
(6) (CONCLUSION from 5 and 2) There are no R-concepts.

It is far from uncontroversial that premise (4) is true. The besires literature is loaded with arguments for and against the possibility of besires. But I'm going to accept the premise: partly because I believe it to be true, and partly it puts my suggestion in a dialectically better position if it's compatible with the impossibility of besires.

The strategy I'll adopt is perhaps surprising. I'll deny premise (3) which says that reactive representations are a kind of besire.

Reactive representations certainly play some the causal roles that beliefs and desires do. But recall that a besire is supposed to be both a belief and a desire. Importantly, as I hope to make clear, it's a type of state that is both a belief with a certain content and desire with certain content. This is in contrast with the idea that there might be a *token* state which is both a belief and a desire.

To see this consider this characterization by Smith of the problem that besires pose for Humeans:

> W]hat Humeans must deny and do deny is simply that agents who are in belief- like states and desire-like states are ever in a single, unitary, kind of state.... And their argument for this claim is really quite simple. It is that it is always at least possible for agents who are in some particular belief-like state not to be in some particular desire-like state; that the two can always be pulled apart, at least modally. This according to Humeans, is why they are distinct existences. (Smith 1994: 119)

The thought here is that if you are in a belief-like state, there is no necessary connection to any desire-like state. If I believe that there are possums around me, and I desire to flee, it's always (logically) possible for me to be in the same belief state but some other desire state: perhaps wanting to cuddle them. But if the belief that possums are around, and the desire to flee, are the very same state—a besire which is both a belief and a desire—then that isn't logically possible. For this state (which is the belief that possums are around) is in part individuated by also being the desire to flee possums. Should I no longer desire to flee possums, I am no longer in the same state: so I no longer have the belief that there are possums around. This violates the Humean doctrine that for any belief, I may combine it with any desire.

This is a doctrine I subscribe to, so this thought should affect me. And it does. If besires are both beliefs and desires in the sense that the belief that P and the desire that Q could both be type identical to a particular state which has belief-like features and desire-like features, then I am persuaded that there are none.

David Lewis' more complicated arguments (Lewis 1988, 1996), which I shan't discuss in detail here, nevertheless depend I think on a similar thought. The rules for evolution of beliefs and desires are different. By "rules" here we mean something like the patterns of evolution which are constitutive of being a belief or a desire. If there is one state which is both a belief and desire in this type sense, the rules of evolution for belief allow it to change in a way which has no effect on its motivational aspects. But that means it's still the same desire—except it isn't, because desires were meant to be identical to states which have the representational features we have evolved away from. Similarly the rules for evolution of desire allow us to evolve away from our desire profile without affecting the world-head aspect of the state. Since the world-head aspects have not changed, we have the same original belief. But by hypothesis our desires have evolved, so we do *not* possess the same desire. But the original belief we possessed was supposed to be identical with this desire, so we cannot have the original belief and fail to possess the desires. So the desired hypothesis in its most flat footed form leads to contradiction. So that's why there can't be any besires, if by besire you mean a type of state that is both a certain belief and a certain desire, and by "is both a certain belief and a certain desire" you mean type identical to both. But are reactive representations besires?

No. Because they aren't type identical to beliefs or desires. Consider the reactive representation 'UGHPOSSUM'. Someone who tokens it perhaps does token, in belief and desire terms, the belief there is a possum and the desire to run away.[18] So 'UGHPOSSUM' plays a role, perhaps, of being the realizer of the belief that there is a possum, and the desire to avoid possums. But it is not type identical to either. Should we evolve in such a way that we no longer take there to be possums around, but are still disposed to run should there be some, then we no longer have the belief that there are possums, we no longer have the RR UGHPOSSUM but we still have the desire to avoid possums, albeit now realized differently. So the RR UGPOSSUM is not identical to the desire to avoid possums, even though it may *realize* it on occasion. Equally, if we remain in a state which covaries with possums, while somehow coming to tolerate them, we may still be in a state which we can call a belief that there are possums around, but we are no longer in UGHPOSSUM, nor do we desire to avoid them. Thus UGHPOSSUM is not identical to the belief that there are possums around, though again it may realize it from time to time.

So a reactive representation is not type identical to any belief, or any desire. A besire, on at least some understandings—the ones that make them problematic—is something that is type identical to both a belief and a desire. Thus RRs are not besires, thus premise 3 of the above argument is false and the argument fails.

Why do I keep saying type identical here, as though I have a contrast in mind? If fact this is optional. I could just say that the RR is not identical to a belief or desire, but that at some times the RR is a realizer of both the belief and desire (or perhaps that some neural state realizes both the RR, the belief and the desire). When desire

[18] Actually I'm not sure we should say that she necessarily tokens the desire to run away. Maybe she only has the desire to run away if the basis for her disposition to run from possums is a more general state that is capable of interacting with various different beliefs.

evolution happens so that we no longer desire to flee possums, but still believe there are some, the token physical state that realized the belief that there were possums and a desire to avoid them is no longer present, but some other physical state exists which realizes the belief that there are possums without realizing the desire to avoid them. The RR too no longer exists. Problem solved.

But type identity talk, as the reader may have guessed, is useful for its contrast with token identity. Exactly how much sense the type token identity distinction makes is controversial. Nothing I say here depends on it being a good distinction, but I take talking about it to be helpful. This is because it explains the sense in which it's easy to see that people think that any kind of talk which is even close to besire talk, is taking seriously the idea of a belief being identical to a desire. And if you think that, you might think the same about RRs, and therefore be puzzled why the standard objections to besires don't apply.

Here then is the thought: you might think that the RR is token-identical to a token of the belief that possums are near, and a token of the desire to avoid possums. Another way of putting this is of course to say that there is just one token state: in virtue of certain features it's right to call it the RR 'UGHPOSSUM'. In virtue of others it's right to call it a token of the belief that there are possums, in virtue of yet other features its right to call it a token of the desire to avoid possums. What happens when that stage changes, so that, for example, it no longer serves the role of generating possum avoidance? The token state at one level of individuation no longer exists. Nor does the desire to avoid possums. But the belief that there are possums does, now realized by a successor token to the original one. Perhaps it will help to think of the relationship between beliefs and desires and RRs as the underlying token states change as something like a counterpart relation: perhaps analogous to Ted Sider's temporal counterpart relation (Sider 2001), except rather than all counterpart relations being of one kind (in Sider's case person counterparts), we can help ourselves to Lewis' notion of the same thing having different counterparts under different counterpart relations. So the initial state is token identical to a belief (that there are possums around), a desire (to avoid possums), and an RR (because the connection between the representation and the behaviour is extremely direct. Suppose the state evolves in such a way as to no longer motivate possum avoidance. The new state bears the *belief-counterpart* relation to the original state (and so we say in ordinary language that the agent has the same belief) but not the desire or RR counterpart. Suppose the state evolves over time in a way that results in the agent no longer behaving as though there is a possum around, but still having the disposition to avoid possums. The new state bears the *desire counterpart relation* to the original state, so we say it is the same desire, but does not bear the belief or RR counterpart relation. Suppose the state, or the agent, evolves in such a way that the agent will act as though there are possums around, still is disposed to avoid them, but that these features are less direct and automatic: the aspect of the agent which registers the possum has to interact cognitively with an interrogable desire to avoid them to produce the behaviour. Then the state bears the belief counterpart relation to the original state, and the desire counterpart relation to the original state, but not the RR counterpart relation. So we say she has the same belief and desire, but not the same RR.

5. Conclusion

In some ways this chapter has been an advertisement for a research programme. What the advertisement does is suggest that there will be substantial explanatory benefits to the idea of an R-concept and its associated RR, and that many of these are of special interest to conceptual engineering and conceptual ethics. If there are such things they may not only explain apparent disagreements where there is not underling disagreement of facts (like the survival case) but also lay bare what is at stake in proposals to reform or change such concepts. The idea unifies puzzles about phenomenal concepts, crypto-evaluative terms, and hate speech. It might explain what is at issue between people who are arguing about crypto-evaluative terms (and perhaps other thick moral terms) and thus when the different participants can or should revise their usage. It provides a tempting explanation of how hate-concepts might be associated with hate speech, and how such concepts differ from merely descriptive ones associated with inflammatory words. It is suggestive of mechanisms where bias can be purely implicit, and inaccessible to an agent honestly interrogating her beliefs and desires. It is equally suggestive of how it might be that linguistic intervention can generate better consequences than just hiding people's racist or otherwise biased attitudes. One way in which this is a very different approach to, say, Haslanger's (2000) view is for her it's much more important what the 'input' conditions are: what it takes descriptively to fall under a concept. Reasons for linguistic intervention in the concepts we have will fall on the side of the social consequences of categorization. I add to this the consequences for the concept possessor; in what possessing such a concept makes one do over and above categorization.

All of this requires that we really have such mental states, but the explanatory merits of it seem to highlight the importance of the empirical project of determining whether we do. But this chapter has at least done some of the groundwork in ruling out two tempting *a priori* objections: that R-concepts function in such an odd way that even if things a bit like them exist, they don't do any of the work that would justify seeing them as a kind of concept, and that they require use to possess besires.

References

Anderson, L., and Lepore, E. 2013. Slurring Words. *Noûs* 47 (1):25–48.

Bermúdez, J. L. 2007. What Is At Stake in the Debate about Nonconceptual Content? *Philosophical Perspectives* 21 (1):55–72.

Burgess, A., and Plunkett, D. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Burgess, A., and Plunkett, D. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102–10.

Braddon-Mitchell, D. 2009. Naturalistic Analysis and the A Priori. In David Braddon-Mitchell and Robert Nola (eds.), *Conceptual Analysis and Philosophical Naturalism*. Cambridge, MA: MIT Press.

Braddon-Mitchell, D., and Jackson, F. 2007. *The Philosophy of Mind and Cognition*. Oxford: Blackwell.

Braddon-Mitchell, D., and Miller, K. MS. Survival in Continuous Degrees.

Camp, E. 2013. Slurring Perspectives. *Analytic Philosophy* 54 (3): 330–49.

Cappelen, H. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Davidson, D. 2004. Paradoxes of Irrationality. *Problems of Rationality*. Oxford: Oxford University Press.
Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.
Egan, A. 2008. Seeing and Believing: Perception, Belief Formation, and the Divided Mind. *Philosophical Studies* 140 (1): 47–63.
Feldman, F. 2010. *What is This Thing Called Happiness?* Oxford: Oxford University Press.
Floridi, L. 2011. A Defence of Constructionism: Philosophy as Conceptual Engineering. *Metaphilosophy* 42 (3):282–304.
Fodor, J. A. 1987. Why There Still Has to be a Language of Thought. *Psychosemantics*. Cambridge, MA: MIT Press.
Haslanger, S. 2000. 'Gender and Race: (What) Are They? (What) Do We Want Them To Be?' *Noûs* 34 (1): 31–55.
Haybron, D. M. 2008. *The Pursuit of Unhappiness: The Elusive Psychology of Well-Being*. Oxford: Oxford University Press.
Hom, C. 2008. The Semantics of Racial Epithets. *Journal of Philosophy* 105: 416–40.
Lewis, D. 1982. Logic for Equivocators. *Noûs* 16 (3):431–41.
Lewis, D. 1988. Desire as Belief. *Mind* 97:323–32.
Lewis, D. 1996. Desire as Belief II. *Mind* 10:303–13.
Margolis, E., and Laurence, S. (eds.) 1999. *Concepts: Core Readings*. Cambridge, MA: MIT Press.
Norby, A. 2014. Against Fragmentation. *Thought: A Journal of Philosophy* 3:30–8.
Nozick, R. 1989. *The Examined Life: Philosophical Meditations*. New York: Simon and Schuster.
Nussbaum, M. C. 2008. Who Is the Happy Warrior? Philosophy Poses Questions to Psychology. *The Journal of Legal Studies* 37:81.
Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.
Pettit, P. 1991. Realism and Response-Dependence. *Mind* 100 (4), new series:587–626.
Price, H. 1989. Defending Desire-as-Belief. *Mind* 98:119–27.
Roskies, A. L. 2008. A New Argument for Nonconceptual Content. *Philosophy and Phenomenological Research* 76:633–59.
Shea, N. 2013. Naturalising Representational Content. *Philosophy Compass* 8 (5):496–509.
Sider, T. 2001. *Four-dimensionalism*. Oxford: Oxford University Press.
Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.
Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: MIT Press.
Väyrynen, P. 2013. *The Lewd, the Rude and the Nasty: A Study of Thick Concepts in Ethics*. Oxford: Oxford University Press.
Zangwill, N. 2008. Besires and the Motivation Debate. *Theoria* 74:50–9.