DOI: 10.1111/phpr.12843

Philosophy and Phenomenological Research

ORIGINAL ARTICLE

Classification procedures as the targets of conceptual engineering

Jennifer Nado

Department of Philosophy, University of Hong Kong, Pokfulam, Hong Kong

Correspondence
Jennifer Nado, Department of Philosophy, University of Hong Kong, Pokfulam, Hong Kong.
Email: jennifernado@gmail.com

Conceptual engineering is a recently (re-)popularized methodological approach which aims at the improvement, rather than the mere analysis, of our current conceptual repertoire. A conceptual analyst might ask whether such-and-so Gettier-style case intuitively counts as a case of knowledge; a conceptual engineer, meanwhile, might ponder the advantages and disadvantages of *altering* our current knowledge-concept to include certain forms of epistemic luck. This sort of concept-tinkering is arguably widespread, and not just within philosophy. Examples of the method range from the 'demotion' of Pluto to the push to remove gender dysphoria from the category of mental disorders. And although the method itself is old, philosophical study *of* the method is still in its infancy, making disciplinary newcomers like experimental philosophy seem downright long in the tooth. Even the name 'conceptual engineering' is new – and even the name prompts philosophical questions about what, exactly, conceptual engineers are really up to.

There's a touch of irony in the fact that 'conceptual engineering' invokes one of the messiest bits of philosophical terminology in current use: 'concept'. On one way of understanding concepts, concepts are psychological entities in the head. But on another way of understanding concepts, concepts are abstracta. Even assuming concepts to be (let's say) psychological entities, there remains extensive disagreement about the basic natures of these entities. Are they prototypes? Atomistic symbols in the language of thought? Recognitional abilities? The 'conceptual engineering' label, taken literally, suggests that the engineer attempts to revise *concepts*. But a literal interpretation invites tricky questions regarding exactly which sort of 'concept' we're meant to be aiming at. Are we trying to alter the contents of a prototype mentally represented in someone's head? Whose - everyone's? Are we trying to alter an abstract Fregean sense? Does that even – pardon the pun – make *sense*?

Of course, we don't have to take the 'conceptual' label literally. An obvious alternative proposal is that conceptual engineers target linguistic entities – expressions of natural language. And here it's a little more straightforward to piece together a reasonable account of what the conceptual engineer does. On the linguistic view, the goal of an engineer is to alter the meaning of a linguistic expression – for instance, to bring it about that the extension of 'marriage' is expanded to include

same-sex couples. But here too, we face puzzles. Does the conceptual engineer really *change* the meaning of 'marriage', or does she simply bring us to realize that 'marriage' included same-sex couples all along?[1] And just how feasible is it, after all, to change the meaning of a linguistic item? As Herman Cappelen (2018) has argued, if we adopt the arguably standard view that meaning is at least in part determined by factors outside the head, our control over the engineering process looks fragile at best.

Have we got any other options? Well, of course we do. We're discussing conceptual engineering here, after all. If targeting linguistic items seems too intractable, and targeting concepts too unclear, why not simply engineer ourselves up a term/concept/category/whatever that characterizes a suitable, tractable, precisely defined target for our engineering efforts? That's exactly what I'll aim to do in this paper. In short, I'll propose that conceptual engineers should take themselves to be in the business of inventing (or perhaps better, discovering) *classification procedures*. Classification procedures, as I'll define them, are essentially 'recipes' for sorting entities – for determining whether a given entity is in or out of the category picked out by the classification procedure. Like recipes, classification procedures are abstract, and may be utilized (often more or less imperfectly) by multiple individuals. Classification procedures are *associated* with linguistic items and with concepts, in the sense that individuals will employ one or another classification procedure to determine whether a given term applies to a given entity, or whether that entity falls under a given concept. But beyond that, the exact relation classification procedures have to linguistic meaning or to conceptual content is - by design - left almost entirely open.

That's the preview, upon which I'll elaborate through the remainder of the paper. In addition to fleshing out the proposal, I'll argue that it has distinct advantages: it evades Cappelen's control problem, for instance, and it links our targets more closely to actual human behavior and reasoning than does a linguistic (or even a conceptual) take on engineering. I'll also explore some of the more interesting upshots of the classification procedures approach. For instance, it implies that all instances of conceptual engineering are 'replacements' rather than 'revisions', and thereby invites a fairly relaxed attitude towards Strawsonian concerns of 'changing the subject'. It also naturally suggests a helpful taxonomy of activities that an engineer might undertake, and illuminates the potentially wobbly boundary between changing e.g. the concept WOMAN and merely changing beliefs *about* women. Ultimately, I'll suggest that the primary strength of the classification procedure approach is that it allows conceptual engineering projects to sidestep potentially tricky issues in the philosophy of language and the philosophy of mind – it is compatible with, and neutral between, all available theories of meaning and all available theories of concepts.

1 | A CLOSER LOOK AT THE COMPETITION: THE LINGUISTIC VIEW

Before setting out the full proposal, let's first consider more carefully the challenges associated with taking either concepts or linguistic items to be the targets of conceptual engineering. I suggest we start with linguistic items, since the surrounding logical terrain is a bit easier to navigate. A linguistic take on the targets of conceptual engineering, as mentioned in the introduction, characterizes the engineer's goal as changing the meaning of a natural language expression. 'Meaning', of course, is itself another one of those notoriously messy philosophical terms. More or less everyone would agree that the *extension* of a term has something to do with meaning, but we certainly can't characterize conceptual engineering merely as 'the process of changing the extension

[1] See e.g. Haslanger (2006).

of a linguistic expression'. On such a view, Fluffy counts as a conceptual engineer of the word 'dog' whenever she births a litter of puppies. A much more reasonable view is that the engineer attempts to change *that which determines* the extension of a linguistic item. Thus, e.g., Herman Cappelen writes that the engineer aims at 'changes in extensions that are driven by changes in intensions' (Cappelen, 2018, p. 62), where an intension is understood as a function from worlds (or something of the like) to extensions.

Cappelen's characterization is quite neutral, and leaves the nature of intensions open. So let's run with it for the moment. The question that now faces us is as follows: how exactly does one go about changing the intension of a linguistic item? The answer to that question will, naturally, depend on our theory of meaning. Suppose we hold that the intension of the word 'marriage' is fixed by a description that speakers mentally associate with 'marriage'. On such a view one changes the meaning of 'marriage' by changing the description that speakers associate with 'marriage', perhaps from 'legally recognized romantic union between a man and a woman' to 'legally recognized romantic union between any set of consenting adults'. Straightforward in principle, though of course depressingly difficult in practice.

Most contemporary philosophers of language, however, would argue that this sort of descriptivist view of meaning is false. By and large, contemporary views on meaning tend to be externalist – they tend to hold, that is, that the meaning of a term is determined (at least in part) by factors external to the mind of the speaker. If that's the right view on meaning, then conceptual engineering starts to look a lot harder. Cappelen (2018) argues this point, noting that externalist views of meaning have the discouraging consequence that meaning change is largely out of our control – especially if one's particular preferred flavor of externalism holds meaning to be determined by occurrences in the past, such as initial baptisms. Worse, though, Cappelen argues that we aren't even in a position to figure out *whether* we have managed to change an expression's meaning. After all, many of the reference-fixing facts may be out of our epistemic reach – as, for instance, facts about initial baptisms often are. What's more, Cappelen argues, the facts that determine meaning change may simply be too complex; we don't currently understand the mechanisms involved, and potentially never will.[2]

I'd argue that there's a related issue lurking here - one that isn't unique to externalist views. On any linguistic approach, the conceptual engineering facts are inconveniently 'hostage' to the metasemantic facts. Even an internalist fan of the linguistic approach should agree that, on her view, the correct metasemantic facts will determine not only *how* we should approach the task of engineering, but also *whether* certain conceptual engineering projects have succeeded. Metasemantic facts will arguably even determine whether a given project even *counts* as conceptual engineering in the first place. Consider one of Carnap's examples: the move to exclude whales and other marine mammals from the category 'fish'. On a descriptivist approach to the semantics of natural kind terms, this looks like a straightforward case of conceptual engineering. Starting in the early 18th century, biologists gradually recognized that cetacea are more akin to land mammals than to fish, and encouraged the scientific community (and eventually the general public) to adopt biological definitions that reflected this. As a result of internalizing these new definitions, the meaning of 'fish' changed.

But on a standard, Kripke/Putnam-style causal-historical metasemantics, the story looks rather different. On that view, 'fish' has always picked out a natural kind, and thus *always* excluded

[2] Several authors have contested Cappelen's claims about the effect of metasemantic externalism on CE – see in particular Pollock (2020), Koch (2021). While these authors argue that we do have sufficient control over meaning change, I'll be arguing instead that meaning change is orthogonal to the conceptual engineering project.

cetacea. 18th century biologists did not re-engineer the meaning of 'fish'; instead, they simply made an empirical discovery about fish, and consequently about the meaning of 'fish'. Perhaps they *thought* of themselves as trying to change the meaning of the word (though of course they didn't have a notion of 'conceptual engineering'); perhaps we should even say that they *were* trying to change the meaning of the word. But on the view of reference under consideration, this would be a nonsensical project from the start – like 'trying' to get Donald Trump to adopt the slogan 'make America great again'.

The linguistic view makes the nature of conceptual engineering hinge on metasemantic issues. That suggests, for one thing, that we cannot feel particularly confident in our conceptual engineering projects until we're confident that we've got our metasemantic theory in good shape. This isn't necessarily an objection – Cappelen even embraces this idea, claiming that 'at the center of any theory of conceptual engineering is a metasemantic theory' (Cappelen, 2018, p. 7). But the following *is* an objection: *why* should our theory of conceptual engineering be constrained by how the metasemantic facts turn out?

I'd suggest that a little reflection will reveal that our conceptual engineering concerns don't ultimately, or fundamentally, track issues of meaning. Let's return to marriage equality, a subject that many of us care deeply about. Suppose that the God of Semantics were to descend from on high and reveal to us that the correct metasemantic theory is such that meaning is wholly determined by factors present at the time of a word's introduction, such that no amount of usage change will ever amount to meaning change. And suppose further that the God of Semantics informs us that, due to the actual facts of its introduction, 'marriage' refers exclusively to partnerships between a man and a woman. Would that matter? I'd suggest it wouldn't, really – we conceptual engineers would continue to attempt to convince others to classify same-sex partnerships together with heterosexual pairings. And we would continue to do so until a sufficient proportion of community members were disposed to act, infer, and speak in a manner that reflects our recommended way of carving up the world.[3]

Depending on which metasemantic theory turns out to be true, the relevant facts about how members of our community act, infer, and speak may come apart from the semantic facts. Our community might legally grant marriage-related rights to same-sex couples; its members might unilaterally treat such partnerships as perfectly equivalent to heterosexual marriages; utterances of phrases such as 'Joe and Omar are married' or 'Sarah just married her girlfriend' might be commonplace and might provoke neither confusion nor objection. Nonetheless, if the reference of 'marriage' turns out to be determined as specified above, those utterances will be false – only male-female pairings will 'really' be marriages. Well - so what? If a community has the dispositions just described, do we really think that we still need to *further* ensure that the *meaning* of 'marriage' has actually changed?

It might be objected that the correct metasemantic theory could not possibly come apart from usage and inference dispositions as dramatically as the case above suggests. One could argue, for instance, that the inability of a metasemantic view to allow meaning change despite dramatic change in usage (as in the case imagined above) is in of itself a *reductio* of the theory. But on the other hand, most philosophers do hold that meaning can come apart from usage quite substantially in any *particular* individual. The inability to allow for error was, after all, one of the death-knells of classic descriptivism. And then there's semantic deference – the idea that

[3] Mark Pinder has informally made a very similar argument during online discussion, though one more closely tied to his view that conceptual engineers target speaker's meaning (for which see Pinder (2021)). He also expresses similar sentiments in Pinder (manuscript).

ordinary speakers defer to the judgment of experts about what belongs in the extension of a term like 'arthritis', thus making reference dependent on social features of the speaker's environment. This would allow, in principle, nearly everyone in a community to use a term in ways contrary to its actual meaning. Plausibly, that's fairly close to being the case with certain terms like 'schizophrenia' (which is very commonly misunderstood to mean having 'multiple personalities').

We could argue further about whether, and to what degree, the cognitive-behavioral dispositions I've gestured at here can come apart from meaning. But why should we? The real point here is that such argumentation is a red herring. When we engage in conceptual engineering, we do so because we care about certain consequences that we hold to result from said engineering – and changes in how our community acts, infers, and so forth are among the most important of these consequences. Exactly *which* cognitive-behavioral dispositions are relevant to the conceptual engineering project is something that we'll return to a bit further on, along with the related question of whether altering such dispositions is the sole ultimate goal of the engineering enterprise. For now, though, the rough-and-ready examples given here should be enough to suggest that metasemantic theory is at best only indirectly relevant to the conceptual engineer's core concerns.

Metasemantic theory in fact *may* turn out to be indirectly relevant to said concerns; to the degree that meaning correlates with the relevant dispositions to act, infer, and so forth, then changing these dispositions will go hand in hand with changing meaning. But, ultimately, even if conceptual engineering *does* inevitably involve changing meaning, it's still not clear that insight into metasemantics is going to give us much of a leg up on figuring out how to engineer successfully. We already know, more or less, how to bring about changes in our community's dispositions to e.g. treat same-sex partnerships equivalently to 'traditional' marriages: we offer up persuasive *arguments* as to why those partnerships ought to be grouped together in a single category. And we don't need much insight into metasemantics to tell us how to give arguments – after all, we've been doing it since Thales.

2 | A CLOSER LOOK AT THE COMPETITION: THE CONCEPTS VIEW

2.1 | The trouble with 'concept'

So what about concepts? As we just noted, what many conceptual engineers seem to really *want* is for their inventions to produce certain inferential and behavioral changes in a target community. This would seem to naturally fit with a view according to which engineers are in the business of altering concepts. It's our community's changing concept of marriage, the argument would go, that has led to a greater tendency to e.g. not immediately infer from a woman's utterance of 'I am married' to the belief that she has a husband, to accept without confusion utterances like 'Joe and Omar are married', and so on. Change the concept, and you change the relevant dispositions.

Again, we've not yet pinned down the exact class of dispositions that are relevant here - presumably, for instance, our dispositions to celebrate marriages with elaborate multi-tier cakes wouldn't be changed much by tinkering with the extension of our concept of marriage. But the general idea that there is a link between changing a concept and changing the sorts of dispositions that interest the conceptual engineer is, I think, intuitive enough. And certainly, a very large proportion of conceptual engineers have found concepts to be the natural choice of 'target'; it would be easier to list engineers that deny the concepts view than to list those that endorse it. A taster menu of a few of the most explicitly concept-based accounts might include Machery (2017), Haslanger (2020),

Isaac (2020), and Pollock (2020); we'll take a closer look at a few of these accounts later in this section.

Before getting into details, though, let's start with some groundwork clarification, as well as some big-picture worries about the concepts approach generally. As noted earlier, the term 'concept' is one of the messiest bits of philosophical vocabulary we possess – just to start, the term is sometimes used to refer to a psychological entity, and sometimes to an abstract one. It is the psychological sense that seems to most naturally fit with the above story about the cognitive-behavioral changes surrounding marriage; it also seems to be the sense that most proponents of concept views on CE have in mind. So let's start there. Let's consider, that is, the proposal that the activity of conceptual engineering involves altering concepts, psychologically understood.

Suppose that my goal is to conceptually engineer the concept DOG. What would that mean, precisely, on the psychological-concept view? The most salient possibility seems to be that it would mean altering the particular mental entities that count as tokens of the concept type DOG, across some relevantly large subset of a target community. What that amounts to depends on what, exactly, those particular mental entities *are*. And here, we run into essentially the same issues that arose in the linguistic case: there's substantive disagreement over what concepts are, but it doesn't seem like resolving those debates is terrifically important to figuring out what conceptual engineers ought to be doing. A psychological-concepts account risks making the conceptual engineering facts inconveniently hostage to psychological facts about the nature of concepts, and it just doesn't seem like conceptual engineering *is* hostage in that way.

Let's spell this out a little. The 'traditional' view of concepts takes them to be mentally represented necessary and sufficient conditions; in the case of DOG, something like MEMBER OF THE SPECIES CANIS LUPIS FAMILIARIS. Prototype views of concepts, meanwhile, take them to be mental representations involving a cluster of typical features, rather than necessary-and-sufficient features; in the case of DOG, something like FURRY, HAS A TAIL, WET NOSE, and so forth. Atomistic views of concepts take them to simply be symbols in the language of thought, with no constituent structure whatsoever. And theory-theories take them to be, as one might guess, mentally represented theories. Of course, there's a whole slew of objections to each of these accounts. The trouble is that most of the debates in this area don't seem to be particularly relevant to the question of how conceptual engineering should be understood or performed.

Consider, for example, the commonly held idea that concepts are the constituents of thoughts. If we hold this to be a constraint on the correct theory of concepts, then our theory of concepts is going to be influenced by certain facts about thought, with downstream implications for conceptual engineering. As a quick example, one such fact about thought is its compositionality. Famously, Fodor and Lepore (1996) argue that the compositionality of thought problematizes prototype views of concepts, because prototypes do not compose (a pet fish is neither a prototypical pet nor a prototypical fish). A fan of a Fodorian take on concepts might perfectly well accept that people *have* mentally represented prototypes, and that this explains why our categorization behavior is impacted by the typicality of our target.[4] But said Fodorian would simply claim that those prototypes are not *themselves* concepts, that possession of any particular element of said prototypes is not required for concept possession, and so forth.

Now, suppose that a philosopher manages to alter the individual mentally represented prototypes of most of a target community – suppose, for instance, that most of us no longer give the HETEROSEXUAL feature as much weight in the prototypes that help us categorize marriages. Can we infer that the philosopher succeeded in re-engineering our concepts (psychologically

[4] For classic discussions of the relevant psychological findings, see Rosch and Mervis (1975) and Rosch (1978).

understood)? Well, we can't be sure yet – it depends on whether the Fodorian stance just outlined is correct. But again, resolving long-standing disputes in the philosophy of mind doesn't intuitively seem to be required before assessing the success of a conceptual engineering project.

The compositionality issue just mentioned is linked to the fact that concepts (in the psychological sense) are standardly taken to be mental representations, and mental representations have semantic content. And even putting questions about the status of prototypes to one side, once semantic content enters the picture we're brought right back round to the sorts of issues that complicate a linguistic approach to conceptual engineering. Many philosophers of mind are inclined to be externalists about mental content – that is, they hold that the contents of our thoughts (and *ipso facto*, the contents of concepts) are at least partially determined by factors outside the head. Indeed, some concept-based accounts of conceptual engineering explicitly endorse externalism about the mental (e.g. Sawyer (2020), Haslanger (2020)). But if the targets of our engineering efforts are externally-determined conceptual contents, then this will have consequences for what it is to engineer a concept. Suppose, taking a cue from Burge (1979), that the content of Joe Average's concept ARTHRITIS is fixed by the experts in his community rather than by some prototype Joe mentally represents. If this is so, then changing Joe Average's concept will require intervening with the experts rather than with Joe. It seems a bit odd that the facts about mental content should affect our engineering strategies in this way – particularly if we are ultimately interested in seeing *Joe's* dispositions change.

At the end of the day, if a substantive externalism about mental content is true, we face much the same situation as noted above in the case of language: a community's dispositions to infer, act, and so forth may come apart fairly dramatically from the contents of the community members' concepts. And outside of an aim to produce changes in the former, it's not clear why we should be interested in changing the latter. As in the case with language, it's entirely possible that changing the relevant dispositions does in fact always bring a change in concept in tow – or even that it necessarily constitutes a change in concept. But it's also (epistemically) possible that this is not true. And our theories of conceptual engineering simply aren't plausibly hostage to these sorts of mental facts.

What about a non-psychological view of concepts as the targets of conceptual engineering? The alternative to viewing concepts as psychological entities is viewing them as abstract objects. Would such an approach fare better? The most prominent approach to concepts as abstracta takes concepts to be Fregean senses (see e.g. Peacocke (1992)). A Fregean sense is a hypothesized component of meaning which determines reference, and which plays a variety of other explanatory roles such as accounting for the difference in cognitive significance between co-referential terms. A sense can be described as a 'mode of presentation' for the referent of a term – a way of 'getting at' the referent. At first glance, this might look fairly promising as a way of specifying the targets of conceptual engineering. The problematic gaps between meaning/content and actual individuals' dispositions that we've been discussing don't seem as evident here; after all, one potential way to understand a 'mode of presentation' is as the way in which an individual *thinks* about the referent. So, perhaps a fan of the 'abstract concepts' approach can claim that a person who thinks 'it is not possible for two men to be married' and a person who thinks 'it is possible for two men to be married' are just grasping marriage via two different modes of presentation. They might then claim that 'changing a concept' amounts to getting people to grasp a different mode of presentation for the category in question.

Nonetheless, taking the Fregean sense route still leaves us tangled up in quite a lot of issues that are less than obviously relevant to the conceptual engineering enterprise. Fregean senses are postulated as a component of meaning – which means that a Fregean conceptual engineer

is likely to remain uncomfortably shackled to the metasemantic facts. We don't want arguments against Fregean senses as components of meaning to doom our account of the nature of conceptual engineering. Similarly for arguments against viewing concepts as abstract entities. Moreover, the notion of 'sense' itself is subject to all sorts of open questions – regarding the ontological nature of senses, their individuation conditions, and so forth. Again, the desiderata for successful resolution of those issues don't necessarily carry over as desiderata for a theory of conceptual engineering. Ultimately, the proposal I'll suggest comes fairly close to the 'Fregean senses' take on conceptual engineering – the targets I propose are abstract entities which might be fairly characterized as something like modes of presentation. But calling these things 'senses' does nothing but muddy the waters here, tying us to various presuppositions and assumptions that are more likely to confuse than illuminate.

2.2 | Concepts and cognitive-behavioral dispositions

Linguistic and conceptual views, I've argued, face similar *prima facie* hurdles. First, 'meaning' and 'concept' are terms enmired in decades of philosophical quagmires which are at best tangential to a theory of conceptual engineering. Second, the plausibility of externalist views of linguistic and mental content threatens to drive a substantial wedge between said contents and the sorts of cognitive-behavioral dispositions that many conceptual engineers seem to really have their sights set on. This second problem might suggest that we ought to simply take these dispositions *themselves* to be the targets of the engineering process. Conceptual engineers, on such a view, would be in the business of altering dispositions to infer, act, and speak. Indeed, we might even claim that those dispositions just *are* concepts, on one of the myriad meanings of 'concept' – ask a psychologist to tell you about concepts, after all, and she's more likely to talk about inferences and behaviors than about semantic content.[5]

One obvious route such a disposition-based account might take would be to adopt an existing internalist view of concepts which centers on such dispositions – likely an inferential or conceptual role view. A few authors, such as Prinzing (2018) and Pollock (2020), have leaned in this direction. But internalist views of concepts may complicate theories of conceptual engineering in their own ways. Here's an example. Thus far, we've been discussing 'cognitive-behavioral dispositions' in a rather intuitive, hand-wavy sort of way. A more specific characterization is clearly needed here - although we might well care about *all* of the various dispositions people in our community have to (e.g.) talk about or infer about or treat same-sex couples in various ways, only *some* such dispositions are intuitively relevant to conceptual engineering. For instance, suppose my uncle believes that same-sex couples can be married, but also believes that such marriages are a sin and verbally abuses every married same-sex couple he encounters. That's unfortunate, but altering those latter dispositions doesn't seem like a job for conceptual engineering, at least if my uncle is genuine in his belief that these are indeed instances of marriage. Conceptual engineering is a particular type of project; it is not simply the practice of changing just *any* sort of belief or behavior.

Now, taking an inferential/conceptual role view of concepts would directly imply a characterization of the relevant subset of dispositions: the target dispositions[6] are the ones that constitute

[5] Thanks to an anonymous reviewer for pushing me on this more psychology-influenced approach to the nature of concepts.

[6] Or whatever other psychological entities are held to cash out 'inferential role' – theories may differ on this.

or determine the content of the target concept. But this answer leads to some notoriously tricky issues. Just *which* inferential dispositions are determinative of a concept's content? If we take 'inferential role' too broadly, then this would seem to imply that very few of us ever possess the same (type) concept – with attendant issues for communication, disagreement, and so forth. But if we plan to carve off some privileged subset of the sum total of inferential dispositions, then we would seem to require something very much like the still-commonly-maligned analytic/synthetic distinction to draw the required division.[7]

Of course, these issues are the direct result of the fact that inferential/conceptual role views of concepts are *semantic* views, aiming to provide an account of a concept's *content*. This is one more reason why I'd argue that a conceptual engineer should avoid hitching her horse to existing accounts of the semantics of concepts, even non-externalist ones. In doing so, they inherit the challenges those theories face. Those challenges may well be surmountable, but it's not clear to me why a conceptual engineer needs to confront them in the first place.

A more promising strategy, to my eyes, would be to try to characterize the sorts of cognitive-behavioral dispositions (or processes, or what have you) that look to be central to CE while remaining neutral on any tie those psychological states might have to meanings/contents. There are only two concept-based accounts that I'm aware of that explicitly take something like this route[8] – those of Edouard Machery (2017) and Manuel Gustavo Isaac (2020). Machery distinguishes between the between the 'cognitive content' of concepts and their 'semantic content' (Machery, 2017, p. 227, fn 13), and aims his methodological proposals at the former; Isaac distinguishes between 'philosophical' (read: semantic) approaches to concepts and 'psychological' ones, and plumps for the psychological approach. Machery holds concepts to consist of belief-like states that are 'retrieved *by default* from long-term memory to play a role in cognition and language-understanding' (Machery, 2017, p. 210, emphasis original), and explicitly notes that "the distinction between what belongs to a concept and what does not is neither semantic nor epistemological; it is through and through psychological" (Machery, 2017, p. 212). Isaac echoes this basic characterization, adding that the relevant bodies of information may take the form of prototypes, exemplars, or theories (that is, he holds concepts to be multiply realizable and characterized by their functional role in cognition).

The notion of 'retrieval by default' for use in cognition, emphasized by both Machery and Isaac, helpfully carves out a set of psychological states that we're meant to be focusing on, without e.g. plunging us into issues surrounding analyticity. Moreover, it meshes well with the usage of 'concept' by psychologists. Nonetheless, it's not clear to me that it's the right way of carving for conceptual engineering purposes. Retrieval 'by default', as Machery characterizes it, requires that the retrieval process exhibit speed, automaticity, and context-independence (Machery, 2017, p. 211). Now, consider the fact that proposals offered up by conceptual engineers almost invariably take

[7] We've got Fodor and Lepore again to thank for this latter argument (Fodor and Lepore, 1993); the former appears in several places, including Fodor (1987) and Block (1986). I'm actually a believer in analyticity, myself – but the point here is that there's no reason why a conceptual engineer should rest her account on such a commitment if she doesn't have to.

[8] An interesting, difficult-to-categorize case here is that of Sarah Sawyer (2018, 2020). Sawyer separates the determinants of linguistic meaning from the determinants of topic/subject matter, and argues that concepts serve the latter rather than the former role. Nonetheless, her view on the process of conceptual engineering remains fundamentally semantic: she claims that conceptual engineering aims at changing linguistic meaning while preserving topic. That is, her account aims to "explain why the representational connection between a term and a topic is not disrupted by intensional and extensional variation of the kind that occurs in conceptual engineering" (Sawyer, 2018, p. 10). As such, concerns we've already discussed will apply. Sawyer also talks of 'conceptions', which consist of beliefs associated with a concept, but she claims that these conceptions do not individuate or constitute concepts – instead, they determine usage, which determines linguistic meaning (Sawyer, 2018, pp 11-12).

the form of definitions, consisting of necessary-and-sufficient conditions for an entity to fall under the target category. But the type of information that gets called up 'by default' when we engage in classificatory cognition is, as noted earlier, arguably not usually in that format – evidence suggests that we're more likely to bring up FLUFFY, BARKS, WET NOSE than MEMBER OF THE SPECIES CANIS LUPIS FAMILIARIS. If conceptual engineering aims to 'change concepts' in the Machery/Isaac sense, then engineers proposing 'classical' definitions may have to effect fairly dramatic changes to our default classificatory processing.

Perhaps that's possible – perhaps sufficient mental training might enable me to, by default, summon up a classical definition when I sort (say) dogs from non-dogs rather than relying on mentally stored information about typical features of dogs. But does conceptual engineering *success* hinge upon whether this occurs? Must use of a definition be 'automatic' before we can say that the job is complete? Not in most cases, I'd think. Perhaps we do want to eventually ingrain an equitable definition of 'marriage' so deeply in our community that it utterly replaces any 'default' presumption of heterosexuality. But if the occasional astronomer has to at times consciously override her life-long tendency to call Pluto a planet, I wouldn't say that the International Astronomical Union needs to add this to the agenda of the next General Assembly. *Mutatis mutandis* for many of the carefully-honed definitions of other technical terms in the sciences, in philosophy, in law, and so forth. What matters is that these definitions are *used*, not that they are used with the sort of automaticity and speed that Machery and Isaac have in mind. In many cases, moreover, we actually *don't* want use of these definitions to be context-independent; it may be more advantageous to switch from a technical, precise definition used in the laboratory or classroom to a looser, less cognitively demanding classificatory process for everyday cognition. Most physicists likely don't (and shouldn't) retrieve THE PRODUCT OF MASS AND ACCELERATION DUE TO GRAVITY when they step on their bathroom scales in the morning.

2.3 | Engineering and Implementation

This brings us to an important distinction, and one last reason why I propose that we avoid characterizing conceptual engineering as the project of altering concepts. When conceptual engineers talk about the conditions for 'successful' engineering, they may have one of two things in mind: *formulating* a superior successor to a given concept, or successfully *implementing* that new successor. The distinction, though easily blurred, has been emphasized in a few recent contributions to the conceptual engineering literature – see for instance Pollock (2019), Jorem (2021), and Koch (2021). Implementing a change to our language or concepts, insofar as this is understood as semantic change, requires changing meaning or content. Implementing a change to our concepts in the Machery/Isaac sense would require altering the 'default' cognitive processes employed in certain types of cognition, presumably over some substantive majority of a target population.

I would argue, however, that the real 'meat' of the engineering process consists in the *formulation*, rather than the implementation, of suitable successors: in finding useful ways of carving up reality, and in constructing definitions to express those useful carvings. Persuading others to adopt those definitions, and thereby changing some relevant subset of their cognitive-behavioral dispositions, is in many cases the ultimate goal – but this latter step is really 'advertising' rather than engineering per se. Moreover, although altering cognitive-behavioral dispositions is clearly one of the primary goals of conceptual engineering, it's arguably not the only goal. Some engineers may simply want to find categories that carve nature at its joints – regardless of what effects such carvings have on our various dispositions. As another example, the goal of conceptual engineering

proposals within the sciences may typically be to improve the predictive-explanatory power, coherence, or simplicity of a theory in which the target term/concept/whatever is embedded. Perhaps this latter can be cashed out in terms of a change in the dispositions of some subset of the members of the relevant scientific community, but such a reduction isn't obviously straightforward.

Along with many other fans of conceptual engineering[9], I'm attracted to an account of conceptual engineering that takes the function, purpose, or role of a term/concept/whatever as central to the engineering enterprise. On this type of view, the term/concept/whatever may be viewed as a *tool*, and our goal in designing such tools is to make them well-suited to the purpose(s) they are intended to serve. In the sciences, for instance, we aim to find ways of classifying that enable greater predictive and explanatory power in our theories. The periodic table of elements groups atoms by reference to how many protons they have, rather than by e.g. how many neutrons they have – because atoms with (say) five protons are much more likely to behave similarly to one another than atoms with five neutrons. Grouping by atomic number thus enables more predictive/explanatory power than the alternatives. Similarly, I'd argue, classifications in social domains can serve various purposes - such as promoting a just and equitable society, for instance. The primary job of a conceptual engineer is to figure out whether (for example) society's goals would be better served by adding additional gender categories to our repertoire – or whether it would be preferable to abandon gender categories altogether.

If this general approach is right, then the majority of the actual cognitive work that goes into conceptual engineering will be at the level of *designing* the tools – that is, in inventing or discovering ways of classifying, and in thinking through how suitable those classifications are for various purposes, functions, and roles we might recruit them for. The process of changing a community's concepts (in either the semantic or the Machery/Isaac sense) is a matter of getting folks to *use* the tools we've designed – again, advertising. An important process, but not an *engineering* process in the normal sense of the word.

Of course, the boundaries blur here. In many cases, the reasoning that goes into determining whether a given classification is well-suited to a role can be directly recruited as 'advertising' – if I reason to myself that a wider definition of marriage is likely to better serve our societal goals, then the most obvious way to 'advertise' is simply to pass this reasoning along to those I aim to convince. But sadly, the world would be a much better place if everyone's minds were easily changed by good arguments. Instead, a depressingly large proportion of our fellow humans seem to be shockingly resistant to being reasoned out of the views they've dug themselves into. All the more reason to retain the label 'conceptual engineering' for the 'invention' phase of the project, I think. Figuring out how to coax entrenched conservatives into accepting marriage equality is largely going to be a psychological and sociological project rather than a philosophical one (more's the pity). Which is not, of course, to say that philosophers can't join that project – philosophers can wear as many disciplinary hats as they'd like.

3 | THE CLASSIFICATION PROCEDURE VIEW

Over the past two sections, I've raised some worries for theories that take word-meanings or concepts to be the targets of the conceptual engineering enterprise. Of course, none of these worries are decisive. The real point here is that they are entirely avoidable. In this section I'll propose that

[9] See e.g. Prinzing (2018), Thomasson (2020), Simion and Kelp (2020).

we just aim at characterizing the targets of conceptual engineering *directly*, without the detour through pre-existing philosophical categories such as 'concept'. Then, if the characterization we engineer up for our target later turns out to *be* the correct characterization of 'concept' or of 'meaning', that's fine. But if it doesn't, that's fine too. Either way, we sidestep a good deal of irrelevant philosophical tangles.

Just to make the argumentative strategy here crystal clear, consider how a conceptual engineer might make a prescriptive suggestion in epistemology. Suppose she argues that reliably formed true beliefs are really the proper aim of scientific inquiry. A detractor might then reply that knowledge is not mere reliably formed true belief, because of e.g. counterexamples like Lehrer's (1990) 'Truetemp' case. The engineer could respond that she doesn't really *care* whether reliably formed true belief is knowledge. Perhaps it is, in which case her claim amounts to the claim that knowledge is the aim of scientific inquiry. And perhaps it isn't, in which case she is claiming that knowledge is *not* the aim of scientific inquiry. She might note, in defense of keeping these questions separate, that certain constraints on a correct theory of knowledge aren't necessarily relevant to characterizing the goal of science. For instance, a successful theory of knowledge ought to imply that most of our commonsense beliefs amount to knowledge; but it's arguably fine if our characterization of the aim of science implies that nearly all everyday beliefs fall short of the scientific standard. Clearly, the aim of science will be something in the *neighborhood* of knowledge – something epistemic, at least. But as a conceptual engineer, she's not obligated to stick with old concepts like KNOWLEDGE in her attempts to characterize this aim.

This is exactly the strategy I'll be following here – I'll aim to engineer a *de novo* characterization of our target, and then let the chips fall where they may as to its relationship with meanings and with concepts. Following the argumentation in the previous sections, I propose that we characterize engineers as producing '*tools*' for classifying – tools which have certain desirable effects when *used*. These effects may include encouraging certain types of inference, speech, or behavior; facilitating prediction and explanation; simplifying theory; and so forth. Rather than identifying these classificatory tools with words or with concepts, I'll characterize them as *methods* – specifically, as what I'll call *classification procedures*.

3.1 | Classification procedures: what they *are*

A *classification procedure* is any procedure that, when followed, allows the user to sort a set of entities into two groups – those 'in' the category delineated by the procedure, and those 'out' of that category. 'Procedure' here is intended in the ordinary English sense; a procedure is a method, a process, a set of steps aimed at achieving a goal. This is admittedly not a particularly precise definition for the term 'procedure'; however, it may be about as good as we can get. The archetypal example of what I have in mind here is an algorithm: an exact, finite series of instructions for performing a computation.[10] An algorithm that allows a computer to pick out all primes from a given set of integers would, for instance, count as a classification procedure. But while all algorithms are procedures, it's plausible that not all procedures are algorithms: algorithms must be unambiguous, they must terminate in a finite number of steps, and so forth. I want to leave open (but not necessarily commit to) the idea that conceptual engineers might sometimes cook up inventions that violate these restrictions, and I'll therefore stick with the more general notion of a 'procedure'.

[10] Definitions of 'algorithm' themselves tend to employ 'procedure' or some similar term (such as 'method' or 'instruction'); thus, though I haven't offered a further analysis of 'procedure', I can at least take comfort in being in good company.

Classification procedures, then, are essentially methods for categorizing. They are abstract entities – a classification procedure itself should be distinguished from any particular use or expression of it. As a specific example, an algorithm for sorting primes (considered as an abstract object) can be distinguished from any particular written program that implements it, or any particular instance of its computation by a specific computer.

If a classification procedure is sufficiently consistent and thorough, it will determinately 'pick out' a function from worlds to sets of entities within that world. This 'corresponding function' will characterize the results of applying the procedure (at the actual world) to each possible world. The output of a procedure's corresponding function when we input a given world is the set of members, at that world, of the category that the classification procedure generates. Multiple procedures may correspond to the same function, just as multiple mathematical expressions may describe the same function (e.g. $f(x) = x$ and $g(x) = 2x/2$)[11]. In such cases, there will be multiple procedures that each 'get at' the same category; in other words, multiple ways of sorting which lead to the same resulting classifications.

The 'corresponding functions' I've just described, you'll note, are structurally identical to intensions. In fact, a classification procedure's corresponding function may turn out to *be* the intension of a linguistic item. But it *need not be*, at least in the sense that some functions from worlds to sets of entities aren't 'attached' to any linguistic item at all. Whether we call such unattached functions 'intensions' (and their output sets 'extensions') is essentially a terminological issue. In any case, let's call classification procedures that unambiguously fix upon such a world-to-set function 'well-defined'.

Some classification procedures will not be well-defined – they will fail to determinately fix upon a corresponding world-to-set function. This can come about in two ways: either the procedure fails to generate any output for a given input, or it generates multiple outputs for a given input. The former sort of failure might result from an incomplete, vague, or open-textured classification procedure.[12] Non-well-defined procedures of this sort are not automatically problematic; after all, there may be pragmatic benefits to classification systems involving vagueness or open texture. The second type of failure – when a procedure generates multiple outputs for an input – reflects an inconsistency in the procedure. In other words, there is at least one world and at least one entity within that world such that the procedure deems it both 'out' and 'in'. This type of failure is, I would think, universally problematic. In any case, for non-well-defined procedures, we can get a rough understanding of the categories they generate by analogy with the well-defined case. An entity at a world is determinately in the procedure's category if it is always part of the output set when the procedure is applied to that world.

To sum up: a classification procedure is an abstract 'recipe' which sorts entities into an 'in'-group and an 'out'-group. Some such procedures – 'well-defined' ones – will determinately pick out an intension-like function from worlds to sets of entities, and multiple procedures may pick out the same function. Non-well-defined procedures will generate either incomplete or inconsistent classifications, and thus will not determinately fix a world-to-set function. Nonetheless, some non-well-defined procedures may be perfectly reasonable tools for classification.

[11] Functions are here being defined in terms of sets of ordered pairs; hence the functions these two expressions describe are numerically identical.

[12] An anonymous reviewer raises the issue of whether or not a classification procedure can be vague. This is a thorny question, one which I don't have strong views on. Perhaps only (e.g) an expression of a procedure in language can be vague, in the sense that it can be indeterminate which (precise) classification procedure it picks out. I'm fairly open to this possibility; if vague procedures don't exist, that's simply one less type of non-well-defined procedure to worry about.

3.2 | Classification procedures vs. linguistic definitions

That's a lot of fairly high-level description, so an example or two will likely be welcome at this point. Fortunately, examples are easy to come by, for any definition or conceptual analysis can be viewed as describing a classification procedure. The definition 'x is a bachelor iff x is an unmarried man', for instance, describes just such a procedure. To follow the procedure, simply place an entity in the 'in' group if that entity is an unmarried man, and place it in the 'out' group if it is not. If we hold the English words 'unmarried' and 'man' to have precise, consistent, non-vague intensions (a big 'if'), then the above phrase will unambiguously express a well-defined classification procedure. The procedure's corresponding function will be the function that, when given a possible world as input, returns the set of unmarried men at that world as output.

Now, if the English word 'bachelor' really *is* correctly defined by 'x is a bachelor iff x is an unmarried man', then the expressed classification procedure's corresponding function will also be the intension of 'bachelor'. If the definition given above *fails* to capture the meaning of 'bachelor' – perhaps the Pope is not really a bachelor – then the described classification procedure will simply pick out a function that happens to be non-identical to the intension-function of 'bachelor'. Which is fine - perhaps a conceptual engineer will come along and argue that the simplicity of the above classification procedure's categorization makes it preferable to the actual classification effected by 'bachelor'.

Relatedly, it is crucial to note that definitions are not themselves the same things as classification procedures. Certainly, a linguistic expression like 'x is a bachelor iff x is an unmarried man' is not a classification procedure – linguistic items are not procedures. If we suppose definitions to be propositions rather than the linguistic expressions that correspond to them, then definitions are still not classification procedures. Procedures are, like propositions, abstract entities. However, procedures and definition-propositions are different breeds of abstracta. Definitions provide the meanings of terms, whereas a procedure need not have anything to do with meaning whatsoever. The utterance "Bring me the red blocks" communicates a classification procedure, but it does not express a definition. A definition-proposition can, of course, *describe* a procedure – the definition of 'bachelor' describes a procedure for sorting the bachelors from the non-bachelors. That same procedure, though, could be communicated to a sufficiently clever non-English speaker via simply pointing to a series of examples, without any reference to the linguistic item 'bachelor' at all. Definitions are a particularly handy way of letting others know what sorting method one has in mind, but they are far from the only option.

This is enough to get us an initial characterization of what it means to take classification procedures as the 'targets' of conceptual engineering. The most natural way to carry out a conceptual engineering project is via offering up a stipulative definition – one which will typically not express the current meaning of the term employed as definiendum. The Pluto-demoting definition of 'planet' mentioned earlier is an example of such a case. This definition, as formulated by the IAU, runs roughly as follows: x is a planet iff x is a sun-orbiting body which is spherical in shape and whose gravitational force has cleared its orbit of debris. Now, a fan of the linguistic approach views this offered definition as reflecting an attempt to alter the meaning of the term 'planet'. A fan of the conceptual approach might view it as reflecting an attempt to alter the content of the concept PLANET. I'm proposing that offering a stipulative definition is a way to *express a classification procedure* that the engineer recommends adopting. As I'll discuss more fully in a moment, this recommendation may involve replacing a classificatory practice *currently associated* with the definiendum – which itself may or may not match the *meaning* of the definiendum. Thus,

the IAU's recommendation was that we use the defined procedure when deciding whether or not something should be called a 'planet', should be treated as a planet, and so forth. In other words, they recommended that we associate this procedure with 'planet'.

3.3 | Classification procedures vs. cognitive-behavioral dispositions

We've noted that classification procedures are abstract entities that are not equivalent to definitions, but may be *expressed via* definitions. Similarly, classification procedures are not equivalent to the cognitive-behavioral dispositions we've been trying to pin down throughout the paper. Just as we can distinguish between a *hammer* (the tool itself) and the *disposition to use a hammer* in a given context (e.g. when one needs to drive in a nail), we can and should distinguish between the *tools conceptual engineers design* and the various *dispositions to use those tools* that members of a given community might possess in various contexts. In this section, I'll suggest that the type of dispositions that are relevant to conceptual engineering are those that depend on dispositions to use a classification procedure in certain contexts. Hence, by recommending usage of a given classification procedure in a given context, we may ultimately aim to modify a range of downstream inferential and behavioral patterns. By recommending that our community employ a new classification procedure when employing or interpreting the term 'married', for instance, we may hope to reduce dispositions to verbally object to utterances like 'Joe and Omar are married'.

As indicated by the example just mentioned, one obvious type of context where a person might hold a disposition to use a classification procedure is when uttering or interpreting a particular linguistic item. More or less by definition, every language-user will have some manner of classification procedure 'associated with' each linguistic item they use. After all, every language-user possesses dispositions to treat certain entities as falling under her linguistic terms[13] – to treat Fluffy as falling under the term 'dog', for instance, and to speak and interpret others' utterances accordingly. For instance, if I have a disposition to treat Fluffy as falling under the term 'dog', I'll be inclined to call her a 'dog' under appropriate circumstances, such as when asked 'what kind of animal is that?'. Whatever process I use to determine whether Fluffy should be so treated – whatever mental calculations are involved, or what have you – this can be viewed as an instance of following a procedure for sorting entities into those that fall under 'dog' and those that do not. In other words, the (concrete) process can always be *expressed* in terms of an (abstract) classification procedure, much in the same way that proponents of a computational theory of mind hold that any instance of mental activity can be characterized as implementing a computation.[14][15] And,

[13] If a user had no dispositions whatsoever to apply or withhold the term, I would hesitate to call her a 'user' of the term at all. Thus, someone who has merely heard her German neighbour use the term 'Weltschmerz' but has no idea how to apply it would not count as a 'user' of 'Weltschmerz'. But if she comes to learn that it is some kind of emotion, she now counts as a minimal user – and, correspondingly, she will have a (non-well-defined?) classification procedure to hand which will at least enable her to deem trees and toasters and mammals 'out'.

[14] Note that due to the looser definition of 'procedure' that I've suggested, the claim that the relevant mental activity can be expressed as a classification procedure is much weaker than the claim that mental activity can be expressed as computation.

[15] Just as with mental computation, the language-user doesn't need to be *consciously* following the set of rules that describes her behavior. Moreover, the classification procedure a language-user associates with a given term may well be inconsistent, or illogical, or massively disjunctive. The question of whether the relevant classification procedure can be vague or incomplete is, as noted in footnote 15, a complicated one. The answer will have consequences for how to characterize the relationship between the particular mental processes of an individual and the abstract classification procedure that best

it's worth re-emphasizing, the procedure that best characterizes the language-user's classification dispositions for 'dog' may have little to do with the actual meaning of 'dog'. It's highly conceivable, for instance, that a child might mistakenly treat foxes as falling under 'dog', and thus make various 'dog'-utterances when in the presence of a fox.

All this goes *mutatis mutandis* for concept-users and concepts: every concept user possesses dispositions to treat certain entities as falling under the concept, and these dispositions can be expressed in terms of the user being disposed to follow a certain classification procedure. If I treat Fluffy as falling under DOG, I'll typically be inclined to make inferences about her according to the beliefs I have about dogs (such as inferring that she is a mammal). And I'll typically be inclined to behave towards her as I behave towards other entities I believe to be dogs (such as running away from her if I am afraid of dogs). And just as with language, if an externalist theory of mind is correct, it may well be that a user's classification dispositions for DOG fail to track the content of DOG. A child who mistakenly treats a fox as falling under DOG may infer that the fox would enjoy a nice belly scratch. What's more, a user's classification dispositions for DOG may come apart from the sort of default information retrieval that Machery and Isaac emphasize. DINOSAUR may serve as a better example than DOG here – my default classificatory response to birds treats them as failing to fall under DINOSAUR, but in more scientific moods I may remind myself that our feathered friends are indeed members of the clade Dinosauria. As this example demonstrates, an individual may use different classification procedures in different contexts-of-use for a single term/concept; some uses may be effortful and conscious, while others may be automatic and tacit. In some cases, multiple classification procedures may be used within mere moments of one another – as in cases of 'self-correction', where an individual's default response is activated and then suppressed or overwritten.

We can distinguish, then, between the meaning of 'dog', the content of DOG, and dispositions to treat various things as falling under 'dog' or DOG; this last is determined by whatever classification-procedure an individual uses to evaluate membership in 'dog' or DOG. As noted at the beginning of this section, we're leaving metasemantic questions open; thus it *may* turn out that (e.g.) the content of DOG is constituted by or determined by an individual's DOG-associated classification procedure. But by characterizing our targets in a semantically neutral way, we can use the notion of a classification procedure to carve out the set of engineering-relevant cognitive-behavioral dispositions without appeal to an analytic/synthetic distinction. That is, we can employ the notion of using a classification procedure distinguish between changes of disposition that result from *mere* belief-changes (such as deciding that same-sex marriage is not sinful), and changes of disposition that concern conceptual engineering.

Let's use arthritis as an example. Suppose the classification procedure I associate with 'arthritis' is that expressed by the definition 'x is a case of arthritis if and only if x is a case of pain and inflammation in a joint or a muscle'. Let's further suppose, for ease of argument, that I associate all of the component terms of that definiens with well-defined classification procedures, so that my 'arthritis' procedure is well-defined. Assume that this procedure has a corresponding function that returns, at every world, all the cases of painfully inflamed muscles and joints.

There are a number of beliefs that I might have about arthritis that I could change while keeping this classification procedure in place – I might come to believe that my neighbor Bill has arthritis

captures those processes; however, I don't think those consequences will affect the basic argument being given here. On either view, we can hold that encouraging users to adopt a preferred classification procedure (albeit perhaps imperfectly or incompletely) is an effective way to alter the sorts of cognitive-behavioral dispositions that are intuitively relevant to the conceptual engineering enterprise.

after chatting with him about his condition, for instance, or I might learn from a book that arthritis is particularly prevalent among the elderly. These changes in belief might bring in tow various changes in inferential or behavioral disposition, such as a tendency to encourage older colleagues to get checked for arthritis. We might view these changes as simply resulting from updating my beliefs about which possible world I inhabit, while keeping my use of classification procedures fixed.

But, of course, I might also change my arthritis beliefs by coming to associate a *different* classification procedure with 'arthritis'. If someone informs me that I am mistaken in my belief that 'arthritis' covers painful inflammation in the muscles, I may change my associated procedure to one expressed by the definition 'x is a case of arthritis if and only if x is a case of pain and inflammation in a joint'. Depending on my current beliefs about the actual world I inhabit, this change in classification procedure may cause a number of other belief changes. Perhaps my previous chat with neighbor Bill involved the claim that he has a painful inflammation in his bicep, so now upon changing my classification procedure I abandon my recently-adopted belief that he has arthritis. And, of course, this latter sort of belief-change can bring in tow various other disposition-changes, as well.

It's this latter sort of change, I claim, that conceptual engineering is particularly concerned with bringing about – changes that result from the adoption of new classification procedures. Note that despite the belief language, the above distinction is perfectly compatible with either externalism or internalism about the contents of belief. None of the above hinges on whether or not my concept ARTHRITIS *in fact* covers inflammation of the muscle (and hence whether my beliefs about, say, Bill's arthritis are true or false). Note also that the distinction doesn't require an analytic/synthetic distinction – to echo Machery's sentiment, the distinction here is being drawn along psychological lines rather than semantic or epistemological ones. It is wholly determined by whether the change in disposition is caused by the implementation of a new classification procedure.

So now we have a nice link between individual use of classification procedures and the intuitive range of dispositions to behave, speak, infer, and so forth that we earlier identified to be central to the goals of the engineering enterprise – a link which is much tighter than it was in the case of meanings or contents. I may treat Joe and Omar as falling under 'married', even though they are not in the extension of that term. I may treat my thigh pain as falling under the concept ARTHRITIS, even though it does not. To do so, I just need to be wrong about what marriage is, or about what arthritis is. That is, I just need to have a classification procedure associated with 'marriage' or with ARTHRITIS which generates a classification that departs from the actual extension of 'marriage' or of ARTHRITIS. It's much harder, however, to see how I could treat e.g. thigh pain as falling under ARTHRITIS without having a classification procedure associated with ARTHRITIS that deems thigh pain 'in' the ARTHRITIS category.

Let's summarize the whole picture. I've characterized the engineer as being primarily in the business of formulating and evaluating abstract classification procedures, which we can conceive of as tools designed to serve a given function or purpose(s) via the 'carving-out' of a useful category. Characterizing our targets as abstract entities rather than psychological states helps to keep separate issues surrounding formulation from issues surrounding implementation; it also fits more comfortably with the notion that a conceptual engineer may have various non-psychological purposes in mind, such as simplifying a scientific theory or delineating a natural kind.

The implementation phase of an engineering project involves convincing others to employ the tools we've designed, and thereby to cut up reality in the manner we have in mind. Typically, one will do this by encouraging others to associate the recommended classification procedure with either a novel or an existing term and/or concept – that is, to use the procedure when determining

what falls under the term/concept. When a conceptual engineer convinces her fellows to associate a new classification procedure with an existing term or concept, this will alter what entities her fellows treat as falling under the term or concept – which will in turn bring alterations in any associated dispositions. If we get people to associate a new classification procedure with 'marriage', we alter which couples people take to fall under 'marriage' - and thus which couples ought to be called 'married', treated as married couples are treated, etc.[16] Viewing our targets as classification procedures, then, can allow us to characterize the entire conceptual engineering process – from formulation to implementation – without the need to take a stance on any substantive questions about the nature of meaning or concepts.

4 | RESIDUAL WORRIES AND CONCLUDING THOUGHTS

There are two residual issues that I want to clarify before wrapping up. First, there are a *lot* of moving parts in the account I've just offered. Words of a language have intensions, which determine their extensions; words are further associated with classification procedures, which in turn are associated with what I've called corresponding functions, which in turn delineate a category, which may or may not contain the same entities as the extension of the word. And classification procedures themselves are expressed by, but not identical to, definitions. Plus we've got concepts, the contents of concepts, and dispositions to treat entities as falling under a term or concept.

You might be wondering why we need all these various layers and distinctions. In part, as already noted, the aim is to keep our targets separate from the linguistic and the conceptual – though, as discussed earlier, some of these distinctions *may* end up redundant – for instance, if classification procedures turn out to be the determinants of meaning, then a procedure's corresponding function will be identical to the intension of the associated term, and the category it defines will be the word's extension. But the goal is to not prejudge this question. Another aim is to keep the abstract notion of a procedure separate from the psychological states involved in use of that procedure, again for reasons already mentioned.

However, you might still be wondering why we need *corresponding functions* in addition to classification procedures. The answer here is that we want the targets of conceptual engineering to be more fine-grained than functions from worlds to sets of entities. Consider the two definitions "x is omnipotent iff x can both lift anything, and can create a stone so heavy that no one can lift it" and "x is a Russellian barber iff x is a barber who shaves all and only those who do not shave themselves". Both definitions are associated with the same function – namely, the one that takes all worlds to the empty set. But they very clearly express two different 'ways of classifying', and those different ways of classifying may have different philosophical uses. Similarly for "x is triangular iff x is a polygon with three internal angles" and "x is trilateral iff x is a polygon with three sides". Conceptual engineers may, in other words, find reasons to invent multiple classification procedures that with necessarily empty (pseudo-)extensions, or multiple classification procedures

[16] This leaves open the possibility that changing someone's classification procedure may not ultimately change some of the dispositions we were interested in. For instance, I might convince my conservative aunt that same-sex couples should be treated as falling under 'married' – and she might then simply turn around and abandon certain beliefs she previously had about marriage, such as 'all married couples should be allowed to adopt', or 'all marriage is holy'. Nonetheless, if I do manage to convince her that male-female and same-sex couples ought to be grouped together in the same category, I must have convinced her that they have *something* important in common. And this, at least, is a step in the right direction. The rest of the work is, I'd argue, not conceptual engineering per se.

with the same (pseudo-)intension – that is, multiple procedures that pick out the same function from worlds to entities. Even when we're considering what we might intuitively think of as a single 'concept' like triangularity, there may be multiple procedures for 'getting at' that 'concept' that we might choose to utilize, and there may be choices to be made regarding which procedure is superior. The procedure expressed by the definition "x is a triangle iff, after counting x's sides, adding 86, then subtracting 86, the result is 3", for instance, could certainly be improved upon.

A second worry that some readers might be harboring has to do with the fact that engineering proposals are offered up via natural-language definitions, and accurately following a classification procedure expressed by a natural-language definition seems to require full knowledge of how to apply its component terms. Wasn't the whole point of the classification procedure proposal to sever conceptual engineering's ties with language? Well, not quite. The point was to sever the tie between theories of metasemantics and theories of conceptual engineering. Nearly all conceptual engineering proposals will, however, be *communicated* via language, with all the potential for misunderstanding that this usually implies.

Attempting to follow a classification procedure is rather like attempting to follow an algorithm for solving a mathematical problem, or a recipe for baking a cake. An individual's attempt to follow the procedure may be less than wholly accurate – she may fail to carry the one when performing addition, or she may mis-measure a cup of flour. Similarly, an individual may fail to perfectly follow a classification procedure if, for instance, she is mistaken about the meanings of certain terms in the definition that expresses the procedure. Supposing, for instance, that the extension of 'marriage' actually does include same-sex partnerships, a person who falsely believes marriage to extend only to heterosexual couples might misapply our previously-discussed 'bachelor' classification procedure and include Elton John within the 'in' group.

It's worth re-emphasizing that language is not the only way to communicate a classification procedure; for some simple procedures I may be able to teach the procedure by example, for instance by sorting a group of small objects into two piles while indicating via ostension certain distinguishing marks that form the basis of the classification. But in most cases, a stipulated definition expressed via language will be a more effective pedagogical tool. If I do communicate my recommended classification procedure via language, I always risk my interlocutor 'latching onto' a slightly different procedure than I intended, due to a difference in our understanding of certain component terms in the classification. In some cases the difference in procedure will be too minor to matter much; in others it will be significant. As a potential example, Sally Haslanger (2000) offers a definition of 'woman' that appeals to the notion of being subordinated on the basis of perceived physical characteristics relating to reproductive role. Two persons adopting this definition may end up associating significantly different classification procedures with 'woman' if they have different understandings of what it is for a person to be 'subordinated'.

Fortunately, there is a remedy for such misunderstandings – more engineering. If I come to find that my interlocutor and I are classifying cases differently despite having agreed to adopt the same definition, that should prompt me to try to produce definitions for each of the substantive component terms in the original definition. Of course, in a sense this only pushes the problem back a step, for the new definitions may contain further terms which we understand in different ways. But in practice, we can reasonably hope to eventually reduce the differences in our classification procedures to a point where they no longer have any noticeable impact. That's at least as good as things usually stand in linguistic communication – arguably, it's likely better.

I'll end by talking a bit about 'Strawson's Challenge'. Many readers will be familiar with this worry, which stems from Strawson's (1963) critique of Carnapian explication. The worry is, in a nutshell, that conceptual engineering threatens to change the meanings of our words or concepts

so much that we 'change the subject'. As a result, we'll no longer be talking about or thinking about the things we were initially concerned with.

It's fairly easy to feel the weight of this objection if we're thinking in terms of, say, a linguistic approach to conceptual engineering. If we change the meaning of 'marriage' too much (or perhaps at all!), we'll simply no longer be talking about marriage. The classification procedure approach, however, gives a slightly different perspective. Conceptual engineering doesn't 'change' a classification procedure – procedures are presumably individuated by their steps, so any change of steps is simply a new procedure entirely. In this sense, every instance of conceptual engineering is either an invention *de novo* or a replacement. In the latter case, the procedure used in such-and-so contexts is replaced by another. There's no such thing as 'revising' a procedure, strictly speaking. I'm inclined to the idea that, as a result, either no engineering projects 'change the subject', or they all do. Questions over whether, post-tinkering, we still have 'the same concept' or something of the like just don't really apply.

The closest analogue to 'changing the subject' on the classification procedure account would be replacing the procedure associated with a linguistic term or concept with a new procedure that is 'too far removed' from the meaning or content of the associated term/concept. But we're not obligated to attach our new inventions to old terms. Suppose we argue for replacing the procedure currently associated with 'marriage' – call it procedure 'A' - with some superior procedure 'B'. We could just as easily argue for abandoning the term 'marriage' and its associated procedure, and then introducing a new term 'shmarriage' which we will associate with procedure 'B' and which we will use in every context in which we previously used 'marriage'. From the perspective of the classification procedure approach, these come to more or less exactly the same thing, excepting e.g. emotional associations that might belong to the term 'marriage' itself. In either case, we are 'changing the subject' to the exact same degree – we are saying 'in these contexts, it is better to think in terms of the category delineated by B than the category delineated by A'. Insofar as the B-category better serves the relevant purposes than the A-category, our claims are justified.

The perspective I've just offered on Strawson's Challenge is reflective of the general perspective I'd urge us to take on conceptual engineering as a whole. Discussions of Strawson's challenge often invoke concerns over concept identity (e.g. Prinzing, 2018; Sawyer, 2018) or over changes in 'what we're talking about' (Thomasson, 2020; Cappelen, 2018). Both of these approaches reflect the tendency to view conceptual engineering in terms of language or in terms of concepts. As I've aimed to show in this paper, viewing conceptual engineering in such terms embroils us, needlessly, in all the complexities of the associated debates in philosophy of language and mind. A successful theory of conceptual engineering doesn't need to adjudicate the criteria for conceptual identity, nor the conditions under which two utterances count as 'saying the same thing' (cf. Cappelen). Conceptual engineering can *use* its own methods to *articulate* its methods – and can, as a result, chart its own path.

REFERENCES

Block, N. (1986). Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy*, 10, 615–678.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy*, 4, 73–121.

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

Fodor, J. (1987). *Psychosemantics*. MIT Press.

Fodor, J. and Lepore, E. (1993). Why meaning (probably) isn't conceptual role. *Philosophical Issues*, 3, 15–35.

Fodor, J. and Lepore, E. (1996). The red herring and the pet fish: why concepts still can't be prototypes. *Cognition*, 58, 253–270.

Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be? *Noûs*, 34(1), 31–55.

Haslanger, S. (2006). What good are our intuitions? Philosophical analysis and social kinds. *Aristotelian Society Supplementary Volume*, 80, 89–118.
Haslanger, S. (2020). How not to change the subject. In T. Marques & A. Wikforss (Eds.), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variation* (pp. 235–259). Oxford University Press.
Isaac, M. G. (2020). How to conceptually engineer conceptual engineering? *Inquiry*. Advance online publication. https://doi.org/10.1080/0020174X.2020.1719881
Jorem, S. (2021). Conceptual engineering and the implementation problem. *Inquiry*, 64(1–2), 186–211.
Koch, S. (2021). The externalist challenge to conceptual engineering. *Synthese*, 198, 327–348.
Lehrer, K. (1990). *Theory of Knowledge*. Boulder: Westview Press.
Machery, E. (2017). *Philosophy Within Its Proper Bounds*. Oxford University Press.
Peacocke, C. (1992). *A Study of Concepts*. Cambridge, MA: MIT Press.
Pinder, M. (2021). Conceptual Engineering, Metasemantic Externalism and Speaker-Meaning. *Mind*, 130, 141–163.
Pinder, M. (manuscript). Conceptual Engineering, Speaker-Meaning and Philosophy.
Pollock, J. (2019). Conceptual Engineering and Semantic Deference. *Studia Philosophica Estonica*, 12, 81–98.
Pollock, J. (2020). Content internalism and conceptual engineering. *Synthese*. Advance online publication. https://doi.org/10.1007/s11229-020-02815-9
Prinzing, M. (2018). The revisionist's rubric: conceptual engineering and the discontinuity objection. *Inquiry*, 61(8), 854–880.
Rosch, E. (1978). Principles of categorization. in E. Rosch & B. Lloyd (eds.), *Cognition and Categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum Associates.
Rosch, E., & Mervis, C. (1975). Family Resemblances: Studies in the Internal Structure of Categories, *Cognitive Psychology*, 7, 573–605.
Sawyer, S. (2018). The importance of concepts. *Proceedings of the Aristotelian Society*, 118(2), 22.
Sawyer, S. (2020). The role of concepts in fixing language. *Canadian Journal of Philosophy*, 50(5), 555–565.
Simion, M., & Kelp, C. (2020). Conceptual Innovation, Function First. *Noûs*, 54 (4), 985–1002.
Strawson, P. F. (1963). Carnap's Views on Conceptual Systems Versus Natural Languages in Analytic Philosophy. In P. A. Schilpp (Ed.), *The Philosophy of Rudolf Carnap* (pp. 503–518). Open Court: La Salle.
Thomasson, A. L. (2020). A pragmatic method for conceptual ethics. In H. Cappelen, D. Plunkett, & A. Burgess (Eds.), *Conceptual Ethics and Conceptual Engineering* (pp. 435–458). Oxford University Press.

How to cite this article: Nado, J. (2021). Classification procedures as the targets of conceptual engineering. *Philos Phenomenol Res*, 1–21. https://doi.org/10.1111/phpr.12843