Conceptual Engineering for Externalists

Jussi Haukioja, NTNU Trondheim, Norway

1. **Introduction**

Conceptual engineering, the project of improving our concepts and other representational devices, has received considerable attention and enthusiasm in recent philosophy. Many of our concepts, it is argued, do not make the distinctions we ideally *should* make, in order to succeed in our (political, philosophical, ethical, practical, and so on) aims. We should therefore strive to revise these concepts. Examples of philosophically interesting concepts that have been argued to stand in need of improvement are those of *truth, belief, race, woman, knowledge,* etc. For simplicity, I will here follow many others in the debates and understand conceptual engineering as primarily consisting in intentionally changing the *intensions* of our *words*.[1] Intensions are here understood as functions from possible worlds to extensions, or less technically as criteria for belonging in extensions. Conceptual engineering thereby involves changing the extensions of our words, not by manufacturing or destroying things, but by changing what it takes to belong in the extensions.

Many theoretically interesting problems connected to conceptual engineering have been pointed out and discussed. My focus here will be on possible problems with combining conceptual engineering and *semantic externalism*, a widely held view regarding how linguistic meaning is determined. Semantic externalists hold that the meanings of our terms (or, at least some of them) are at least partly dependent on external matters of fact. Semantic externalism comes in many flavours—my main focus here will be on the two most discussed and widely accepted externalist views, natural kind externalism and social externalism. Natural kind externalism is typically motivated by thought experiments such as Putnam's Twin Earth (Putnam 1975), and holds that the intensions of natural kind terms, such as "water," are partly determined by features of our natural environment, such as the chemical structure of the local watery stuff (that is, the tasteless, colorless substance predominately found in lakes, rivers, taps, and so on). Social externalism, on the other hand, holds that the intensions of many (possibly all) terms are partly determined by facts concerning other speakers (e.g., Burge 1979). One widely held social externalist view, which we will come back to below, claims that the meanings (and thereby intensions) of some terms are determined by *experts* who can make the appropriate distinctions, while the rest of us use the relevant terms with the same meaning as the experts, because we semantically *defer* to these experts (Putnam 1975).

2. **The Problem**

It is not hard to see how a potential tension between conceptual engineering and semantic externalism arises. As noted above, conceptual engineering involves intentionally changing the intensions of our terms. According to semantic externalism, on the other hand, the intensions of our terms can depend on external matters of fact such as chemical structures,

1 I am understanding conceptual engineering to operate on semantic meaning. This assumption is widespread, but not universally accepted: see Pinder (2021) for a defense of conceptual engineering as primarily concerned with speaker meaning.

and/or the beliefs and linguistic behavior of experts. Typically, we have little or no control over such facts:

> […] effecting conceptual change looks comparatively easy from an internalist perspective. We can revise, eliminate, or replace our concepts without worrying about what the experts are up to, or what happens to be coming out of our taps. From the externalist's point of view, however, conceptual revolution takes a village, or a long trip to Twin Earth. (Burgess and Plunkett 2013, 1096)

Steffen Koch spells out the problem as follows:

> (1) SE [semantic externalism] is true about many terms in our language, and in particular those terms typically in the focus of practitioners of CE [conceptual engineering].
> (2) If SE is true about a given term *t*, then it is not within our control to change the meaning of *t*.
> (3) If it is not within our control to change the meaning of *t*, CE is not applicable to *t*.
> (4) Therefore, CE is not applicable to many terms of our language, and in particular it is not applicable to those terms typically in the focus of practitioners of CE. (Koch 2021, 330–331)

Note, however, that at least some *social* externalist views appear to be relatively unproblematic, with respect to conceptual engineering. In particular, the kind of view mentioned above, which holds that the intension of term *t* is determined by the relevant experts' usage (to which non-experts defer), does not pose any special problems for conceptual engineering. Depending on how the experts' usage determines the intension, we get two main kinds of case. In the first, the intension is determined by the properties/descriptions/definitions associated with *t* by the experts. When this is the case, conceptual engineering may of course be challenging for various pragmatic or social reasons, but there is no deep conceptual problem: if the experts agree to change the definition (etc.) that they associate with the term, while the rest of us go on deferring to them, the intension of the term has changed. Arguably, this is exactly what happened when the International Astronomical Union changed the definition of "planet" in 2005. If, on the other hand, the intension of a term is determined by the experts' causal interactions with the kind/phenomenon in question (as it arguably is in Putnam's influential examples of "elm" and "beech"), social externalism does not cause any *extra* problems: whatever difficulties there are, in engineering the meanings of such terms, stem from them being natural kind terms.[2] Accordingly, my main focus below will be on natural kind externalism.

Koch's solution, as mine, is to reject (2). I will discuss Koch's view, as well as present my own, in the next section. But it should be noted that neither (1) nor (3) is obviously true: one could also react to the problem by denying one of them. As for (1), none of the examples mentioned in the introduction are obvious examples of natural kind concepts, although some would hold that, e.g., *knowledge* is a natural kind. However, there's no obvious reason

2 Another kind of social externalism might hold that the intensions of some or all terms are determined socially, but without deference to a particular set of experts. It might, for example, be held that individual speakers defer to how the *majority* of other (competent) speakers in their linguistic community use said terms. Such a view would, for example, seem to fit well with Burge's (1979) discussion of terms like "sofa," although Burge does not explicitly commit himself to it. I assume that such social externalist views would not cause principled problems for conceptual engineering—in the case of such terms, conceptual engineering would merely require changing the speech patterns of the majority of speakers in a community—but I will not discuss this issue in detail here.

why some of our natural kind concepts might not stand in need of improvement: denying (1) would seriously limit the scope of conceptual engineering.[3] Cappelen (2018) is plausibly read as denying (3). I will not discuss his positive view here (but I will, in Section 4, comment on his objection to the kind of view I propose below)—here it is enough to note that his view, too, is unduly pessimistic about the scope and prospects of conceptual engineering, if (2) can be rejected.

3. **Rejecting (2): Semantic Externalism and Meaning Control**

3.1. Koch's Proposal

Koch's solution to the problem starts with the observation that all main variants of semantic externalism already allow for reference change (where this is a result of a change in an externally determined intension, rather than merely a non-semantic change in the world, causing changes in extensions while the relevant intension remains unchanged). This is apparent, for example, in discussions of so-called *slow switching* cases, where a speaker is transported to a new environment containing a natural kind superficially similar to a kind the speaker was previously familiar with, but with a different underlying structure (Burge 1988). It is generally assumed by externalists, for example, that although an Earthling's early tokens of "water" after arrival on Twin Earth would only denote H_2O, were the speaker to remain on Twin Earth and keep calling XYZ "water," eventually her tokens of "water" would change their meaning, and their extension would then include XYZ. These two cases are discussed in some detail by Koch, in his thought experiments of *Young-Mary* and *Old-Mary*, respectively (Koch 2021, 336–337).

Different externalist theories would account for such changes in different ways (cf. Evans 1973; Devitt 1981). For example, according to Evans's theory, which Koch chooses as his illustrative example, a natural kind term such as "water" refers, roughly, to the substance that is the causal source of the body of information that the speaker associates with the term. For an Earthian speaker who has recently been transported to Twin Earth, H_2O is still the main causal source of the information she associates with "water," but after a sufficient time, XYZ will have taken its place, as now most of the information the speaker associates with "water" will have XYZ as its causal source. When that has happened, the meaning of "water," as used by the speaker, has changed. The details of the explanation are not crucial here—what matters for Koch's view is that we *already* think semantic externalism (and natural kind externalism in particular) is consistent with a term's intension changing over time. Provided we have some account of when and how intensions change, what would then stop us from effecting such changes intentionally?

Externalism is then, Koch argues, compatible with what he calls *collective long-range control:* by collectively adopting new ways of speaking about (*e.g.*) natural kinds, we can intentionally bring about meaning change over time (assuming standard externalist views of meaning change are at least roughly along the right lines):

3 But see Haslanger (2006) for the view that something like natural kind externalism applies much more widely than often assumed, in particular that it applies to the social kind terms often focused on in discussions of conceptual engineering.

> Many people start using the term in question *as if it had the new reference*; eventually, this will add pieces to the body of information we associate with the term that have the new object or kind as their causal source. […] Thus, little by little, the term will shift from the old reference to the new one [...]" (Koch 2021, p. 343).[4]

I fully agree with Koch that collective decisions regarding language use can result in intentional meaning change, even if externalism is true of the relevant expressions. However, I disagree with Koch's explanation of *how* such collective decisions can change meanings. In the next section, I will argue that meaning change is in fact, in a sense, easier to accomplish than Koch would allow for, even of externalism is true.

Moreover, I am not convinced that slow switching cases provide us with a good model for explaining intentional meaning change. In slow switching cases, there is by hypothesis no change in the communicative behavior of the Earth-to-Twin-Earth traveler: she continues to apply the term in the same way as before, based on how the situation appears to her. Yet, we are inclined to say that at some point the truth value of her utterances of, say, "there is water in that lake" (pointing to a lake on Twin Earth), will change. The meaning change is not caused by changes in how the speaker is disposed to apply the term, but rather by changes in the environment, of which the relevant speakers are moreover typically assumed to be unaware of. In the kinds of conceptual engineering projects that Koch envisages, by contrast, the environment remains unchanged in the relevant respects: the supposed change in meaning is a result of changes in how the speakers apply the term in question, based on how the situation appears to them. Given this asymmetry, it is not at all obvious that the rate at which the meaning change occurs is similar in in the two cases. In the next section, I will argue that there is good reason to think that the two kinds of situation are crucially different.

3.2. An Easier Way to Reject (2)

Let us start with a thought experiment. Suppose that, sometime in the future, humans discover Twin Earth, which is just as Putnam (1975) famously imagined it to be. Suppose, moreover, that the distance between Earth and Twin Earth is manageable for the technology then available, and we begin frequent travel between Earth and Twin Earth. The chemical difference between the planets is by then well known, of course, and at first speakers take great care to keep track of which planet they are on, and call the liquid they are dealing with either "water," or "twin water," accordingly. However, as the interplanetary travel goes on, this gradually becomes perceived as an unnecessary cognitive burden on speakers – the difference has no impact on their daily lives, after all. And the Twin Earthlings will of course go on calling water "twin water" and twin water "water," just as meticulously, making things even more confusing. Sooner or later, the speakers (of both English and Twin English) decide that life would be a lot easier if everyone just used "water" to talk about watery stuff – *any* clear, odorless, thirst-quenching liquid that fills lakes and rivers, comes out of taps, and so on. This suggestion gains wide acceptance, the populations of the two planets are informed, and everyone conforms to the new usage. ("H_2O" and "XYZ," or

4 Based on the quotation, it might seem as if Koch takes the reference shift to be gradual. However, I think it is charitable to interpret him as claiming the reference shift to be instantaneous: what is gradual is, rather, the process of the *preconditions* of reference shift gradually building up. I am grateful to an anonymous reviewer for pointing this out to me.

some newly introduced terms, are then used in contexts where the chemical composition does make a difference.)

In this imagined scenario, all speakers (of both English and Twin English) switch from applying "water" on the basis of (assumed) sharing of chemical structure with the watery stuff on their respective home planets, to applying it merely on the basis of manifest properties.[5] This fits Koch's description of how we effect "long-term collective control" over the meanings of our terms: speakers "start using the term in question *as if it had the new reference.*" On his view, then, the extension change would take place only after a substantial delay (when the new usage has become the main causal source of information, or the new usage has been in place long enough for multiple grounding to have taken place – the details will depend on our preferred externalist theory of reference change). But can this be right? Remember that the change in the speakers' speech patterns is imagined to be more or less instantaneous: all Earthlings and Twin Earthlings decide to use "water" to refer to all watery stuff, interpret each others' use of the term in the same way, and communicate perfectly using the term. Yet, according to Koch, we should say that the Earthlings' "water" continues to refer only to H_2O, and the Twin Earthlings' "water" to XYZ, for a substantial amount of time, and that speakers utter systematic falsehoods in a substantial range of cases, until at some point in the future the semantic facts click in place. Moreover, when the semantic facts *do* click in place, the only thing that really changes is the truth values of the sentences uttered by the speakers: all the changes in the speakers' communicative behavior took place long before this.

This should strike us as an odd consequence. According to our ordinary practice of assigning truth values, we should surely say that the reference change takes place as soon as the new usage is stable and internalized, whatever this precisely amounts to, just as we say that the meaning of "planet" changed (more or less) instantaneously in 2005, when the IAU decided to change the definition (assuming that the rest of us in fact do defer to the IAU on this matter). But the crucial question is: can we really say this without abandoning semantic externalism? I think we can. In the rest of this section, I will explain how, and in doing so also clarify the relevant difference between slow switching cases and the kind of intentional meaning control consistent with externalism.

A semantic externalist is committed to saying that the meanings of (at least some) terms are partly determined by external matters of fact. Given a term *t*, the meaning (and thereby intension) of which is externally determined, we should separate two questions:

(1) *What kind of* external matters of fact are relevant for determining the intension of *t,* and *how*?

(2) What *are* the relevant facts pointed at, in our answer to question (1)?

For example, if we accept anything roughly like the Putnamian view of "water," the answer to (1) is: the chemical constitution of the local watery stuff matters; sharing this is necessary

5 It might be objected that the change imagined here is so dramatic that it amounts to a change of *topic* rather than a meaning change that is consistent with speakers still discussing the same topic. The question of topic continuity is another contested issue connected to conceptual engineering (see, *e.g.,* Cappelen 2018; Sawyer 2018). A discussion of topic continuity falls outside the scope of the present paper: if it turns out that there is no topic continuity in the case imagined here, a structurally similar thought experiment could be formulated, where the change in meaning is less dramatic (and consistent with topic continuity).

and sufficient for belonging in the extension of "water."[6] The answer to (2), on the other hand, is: the chemical constitution of the local watery stuff is H_2O. Something structurally similar will hold for all natural kind terms, if Putnam is to be believed.

Typically, we have very little or no control over the answers to question 2. There is little we can do about the molecular structure of the watery stuff on Earth. It is precisely this lack of control that motivates doubts about combining conceptual engineering and semantic externalism. However, this leaves open the possibility that we *may* have control over the answers to question 1: we may have control over *which* (and even *whether*) external matters of fact are relevant for determining the intension of a given term, and how such external matters of fact affect the intension. If our pre-theoretical judgments regarding correct assignment of content and truth value are to be trusted, my thought experiment illustrates that we, at least in some imaginable cases, *do* have such control: the speakers in the thought experiment collectively changed the answer to question (1) to (roughly): "no external matters are relevant," thus removing the relevance of any answer to question (2), for "water."

Note also that this is not at all what happens in slow switching cases! In slow switching cases, the answer to (1) remains unchanged: what changes is the answer to question (2). The relevant changes in slow switching cases are *by hypothesis* changes that are, at least ordinarily, beyond our control, and that can happen without the relevant speakers becoming aware of them. Once we notice that the answers to question (1) are just as relevant for determining the intensions of our terms, and that there is no *prima facie* reason to think we lack control over these, the tension between semantic externalism and conceptual engineering should begin to seem much less serious.

Here is another way to put the point. Semantic externalism claims that the supervenience basis of meaning includes external factors, such as the actions of other speakers, facts about underlying natures, and so on. Typically, we have little or no control over these external factors. But we may, nonetheless, have control over *what kinds of facts* are included in the supervenience basis that determines the meaning of a given term. Exactly what determines the supervenience basis for a given term is an enormously complex issue that I cannot hope to settle here, but the following rough sketch seems plausible to me, both when applied to the thought experiment above, and when considered in the abstract. The supervenience basis for a given term—which factors enter into determining its meaning—is dependent on (relatively) stable patterns of use, or perhaps stable patterns in dispositions to use, the term in question. What makes it the case that a given term has an externally determined meaning—and thereby that there exists a positive answer to question (1) for that term—is that the speakers using the term are disposed to *treat* some external facts as relevant when evaluating whether something falls under the extension of the term. If Putnam is right, the meaning of "water" is partly dependent on the fact that our local watery stuff consists of H_2O. What makes it the case that it is *this* external fact which partly determines the meaning, rather than some other external fact, or no external fact at all, is the fact that

6 It is not obvious that this Putnamian view is correct: some recent empirical evidence suggests that ordinary speakers take sharing the chemical constitution of the local watery stuff necessary, but *not sufficient* for belonging in the extension of "water" (cf. Haukioja, Nyquist & Jylkkä 2021). Such details concern, however, only the precise contents of the correct answers to (1), and do not affect the main point of this paper.

ordinary speakers (or, perhaps, expert speakers that ordinary speakers are disposed to defer to) are disposed to take information about the underlying nature of the local watery stuff as *relevant* for evaluating the correctness of the use of "water." For many other terms, such as "bachelor," we do not have similar dispositions: we would not take information about underlying properties of local bachelors to be relevant for evaluating the correctness of using "bachelor."[7]

This kind of a view—which can be fleshed out in more systematic detail by dispositionalist theories in *meta-metasemantics*, such as (Cohnitz & Haukioja 2013) and (Johnson & Nado 2014)—suggests that answers to question (1), for terms with externalist metasemantics, are at least in principle in our control. The kind of coordinated action described by Koch *can* change the meanings of our terms even if semantic externalism is true—in fact, it can change meanings much faster than Koch himself is prepared to allow.[8] There may be all kinds of *practical* difficulties in getting people to change the ways they speak, but a systematic change in how we are disposed to speak and interpret others can change meaning, and semantic externalism does not pose a principled obstacle.

4. **"This Is Not Externalism!"**

Herman Cappelen (2018) considers, and dismisses, a position much like the one I sketched in the previous section. His main target is Peter Ludlow, who argues that "it is within our control to defer to others on elements of the meaning of our words [...] and it is also within our control to be receptive to discoveries about the underlying physical structure of the things we refer to" (Ludlow 2014, 84). Cappelen replies:

> Here is a way to understand Ludlow's position: […] what makes it the case that externalism is true is that we, in a particular conversational setting, decide that it is. According to Ludlow, if a form of externalism is true for a conversation at a time […], that is because the conversational participants […] want it to be true at that time – because they choose to defer to whatever external factors the relevant form of externalism appeals to.
>
> […]
>
> This, however, is not externalism. Externalism as I have understood it […] is not the view that conversational partners at any point in time can just decide that externalist constraints on semantics don't apply. (Cappelen 2018, 166–167)

I am not going to defend Ludlow's theory, specifically, against Cappelen's charge here (but I do believe Cappelen's characterization of Ludlow's view to be uncharitable). When it comes to the view I've sketched—which also claims that it is in a real sense within our

7 There are some who would apparently disagree (see Biggs & Dosanjh 2021), but a discussion of their view will have to wait for another occasion.

8 A dispositionalist view can also provide an explanation of *when* meaning change occurs in slow switching cases, in terms of the relevant speakers' total dispositional states. The reason why Koch's Young-Mary, recently transported from Earth to Twin Earth and unaware of the chemical differences, refers to H_2O with her "water" is, arguably, that were she to learn of the differences, she would *retract* her application of "water" to the watery stuff on Twin Earth. The reason Old-Mary, on the other hand, refers to XYZ is that she would *not* so retract her usage. The meaning change occurs (possibly in a gradual fashion), as her dispositions to retract change. For a more detailed and systematic explanation of meaning change along these lines, see Cohnitz & Haukioja, forthcoming. For a similar account, see Nyquist (2020).

control whether we defer to others, or are receptive to empirically discoverable factors in assigning meanings to our terms—it should be obvious that Cappelen's criticisms are off the mark. Meanings, including whether and how they are dependent on external factors, are determined by systematic and relatively stable patterns of dispositions among language users. These cannot be changed at a whim: the relevant dispositions are relatively automatic and not based on conscious deliberation. We have reason to expect that such dispositions are difficult to change. But, unlike Cappelen seems to assume, we are not faced with a choice between no control at all on the one hand, and freely chosen (meta)semantics on the other.

The interesting question is whether semantic externalism presents a *principled* obstacle for meaning control and conceptual engineering. I've argued that it doesn't. It may well be that successful conceptual engineering is hard to carry out, but that was to be expected. I hold that my thought experiment about frequent travel between Earth and Twin Earth, though no doubt fanciful in its content, nonetheless presents a case where speakers would have a widely shared practical motivation for changing the meaning of "water." Given the motivation, I think it would be realistic to expect that they would engage, and succeed, in the kind of coordinated action required for changing the meaning. That a term has externalist metasemantics may affect *how* its meaning is to be intentionally changed, but it does not preclude *that* we can intentionally change its meaning.[9]

References

Biggs, Stephen & Ranpal Dosanjh (2021). "Pervasive Externalism". In Biggs, S. and Geirsson, H. (eds.), *The Routledge Handbook of Linguistic Reference,* 309-323. New York: Routledge.

Burge, Tyler (1979). "Individualism and the Mental". *Midwest Studies in Philosophy* 4: 73-122.

Burge, Tyler (1988). "Individualism and Self-Knowledge". *Journal of Philosophy* 85: 649-663.

Burgess, Alexis, & Plunkett, David (2013). "Conceptual ethics I". *Philosophy Compass*, 8, 1091–1101.

Cappelen, Herman (2018). *Fixing language.* Oxford: Oxford University Press.

Cohnitz, Daniel & Jussi Haukioja (2013). "Meta-Externalism vs. Meta-Internalism in the Study of Reference", *Australasian Journal of Philosophy* 91, 475-500.

Cohnitz, Daniel & Jussi Haukioja (forthcoming). *Foundations for Metasemantics*. Oxford University Press.

Devitt, Michael (1981). *Designation*. New York: Columbia University Press.

Evans, Gareth (1973). "The Causal Theory of Names". *Proceedings of the Aristotelian Society Supplementary Volume* 47: 187–225.

9 I am very grateful to Daniel Cohnitz, Jeske Toorman, and members of the NTNU reading group on conceptual engineering, for helpful comments on previous drafts of this paper.

Haukioja, Jussi, Mons Nyquist & Jussi Jylkkä (2021). “Reports from Twin Earth: Both Deep Structure and Appearance Determine the Reference of Natural Kind Terms”. *Mind & Language* 36, 377-403.

Haslanger, Sally (2006). “What Good are Our Intuitions?”. *Proceedings of the Aristotelian Society Supplementary Volume* 80: 89–118.

Johnson, Michael & Jennifer Nado (2014). “Moderate intuitionism: A Metasemantic account”. In A. Booth & D. Rowbottom (eds.), *Intuitions,* 68-90. Oxford: Oxford University Press

Koch, Steffen (2021). “The Externalist Challenge to Conceptual Engineering”. *Synthese* 198: 327-348.

Ludlow, Peter (2014). *Living Words*. Oxford: Oxford University Press.

Nyquist, Mons (2020). “On Complete Information Dispositionalism”. *Philosophia* 48: 1915-1938.

Pinder, Mark (2021). “Conceptual Engineering, Metasemantic Externalism, and Speaker-Meaning”. *Mind* 130: 141-163.

Putnam, Hilary (1975). “The Meaning of ‘Meaning’”. In *Philosophical papers, vol. 2: Mind, Language and Reality*, 215-271. Cambridge: Cambridge University Press.

Sawyer, Sarah (2018). “The importance of Concepts”. *Proceedings of the Aristotelian Society* 118:127-147.