*Pre-print – please cite the published version (Synthese 2018)*

**The Externalist Challenge to Conceptual Engineering**

Steffen Koch

Ruhr-University Bochum

[steffen.koch@rub.de](mailto:steffen.koch@rub.de)

**Abstract** There is currently a growing interest in ‘conceptual engineering’. Unlike conceptual analysis, conceptual engineering does not seek to describe the actual meanings of our concepts, but to improve on these meanings. For this method to show the results that its practitioners typically aim for, being able to change meanings seems to be a crucial presupposition. However, certain branches of semantic externalism raise doubts about whether this presupposition can be met. To the extent that word meanings are determined by external factors such as causal histories or microphysical structures, it seems that they cannot be changed intentionally. This paper is an extended discussion of this ‘externalist challenge’. Pace Herman Cappelen’s recent take on this issue, it argues that the challenge is real, i.e., that the viability of conceptual engineering really depends on our ability to bring about meaning change. Furthermore, it argues that, contrary to first appearance, causal theories of reference do allow for a sufficient degree of meaning control. To this purpose, it argues that there is a sense of what is called ‘collective long-range control’, and that popular versions of the causal theory of reference imply that people have this kind of control over word meanings.

**Keywords** conceptual engineering; semantic externalism; causal theory of reference; philosophical methods; metaphilosophy

**1 Introduction**

Philosophy is currently witnessing a growing interest in what has become known as ‘conceptual engineering’. Unlike conceptual analysis, conceptual engineering does not aim to identify the content that our current concepts *do* have, but the content which these concepts *should* have. Conceptual engineering thus aims to revise our current conceptual apparatus according to certain normative standards. Many philosophers believe this to be an underexplored but very fruitful approach to doing philosophy. In this spirit, Herman Cappelen writes that the “[a]ssessment and improvement of concepts is at the core of philosophical practice, no matter what the topic” (Cappelen 2018, 47-8), Sally Haslanger urges us to ask whether our actual concepts are “effective tools to accomplish our (legitimate) purposes” and, if not, to find concepts that “would serve these purposes better” (Haslanger 2000, 33), and Matti Eklund proclaims that “[p]hilosophy should rather be thought of as *conceptual engineering”*,by which he means that *“*[w]e should consider what the best concepts to employ are like” (Eklund 2015, 2).

Conceptual engineering comes in different varieties. In some instances, a concept is modified or exchanged while the word expressing it is held fixed. In other instances, the ‘engineered’ concept is expressed by altogether new vocabulary. Cases of the former type, call them ‘conceptual re-engineering’, have generated particular interest in the recent debate (Haslanger 2000, Brun 2015, Cappelen 2018). What is distinctive of conceptual engineering *qua* re-engineering (‘CE’, from now on) is that it requires the relevant lexical items to change their meanings. For this reason, a theory of this method will have to consider the features that determine linguistic meaning. If a given lexical item has its meaning determined by some condition *c*, then what conceptual engineers ought to do is to change *c*. But what are the kinds of conditions, or features, that determine the meanings of lexical items? Answering this question is the central concern of metasemantics. Metasemantic theories can broadly be divided into internalist and externalist ones. According to internalists, the meanings of lexical items are entirely determined by the mental states of their users (Chomsky 2000, Farkas 2008, Mendola 2008). Externalists, on the other hand, argue that at least some lexical items have their meanings partly determined by facts external to the mind of their users, such as the causal history of their use (Kripke 1980), the beliefs of experts (Burge 1979), or the environment we as a matter of fact live in (Putnam 1973).

At least at first glance, it is much easier to make sense of CE on metasemantic internalism. On internalism, what a speaker ought to do to change the meaning of a term *t* is to change her attitudes towards *t* – her beliefs about what *t* means and/or her intentions to use *t* to refer to something. On externalism, on the other hand, it seems that changing one’s attitudes towards *t* will often not be sufficient to change the meaning of *t*, for *t*’s meaning is at least partly determined by things outside of our cognitive reach. For this reason, it seems that internalists typically allow for a much greater degree of ‘meaning control’ (Ludlow 2014, 83) than externalists do. In light of how appealing at least some variants of externalism are to many philosophers, this raises fundamental questions about the activity of CE, such as these: Is our optimism about being able to change significant aspects of our language well founded? If not, then doesn’t this support a skeptical stance on CE as a philosophical method? Call this the ‘externalist challenge’.[[1]](#footnote-1)

This paper aims to give a thorough analysis of the externalist challenge, to outline a number of theoretical options of how to deal with it, to review some of the attempts at dealing with it that have been proposed in the literature, and, finally, to present a novel solution. Pace Ludlow (2014), I argue that semantic externalism does in fact impose serious restrictions on the amount of control we have over word meanings; pace Cappelen (2018), I argue that a lack of control over word meanings does in fact stand in the way of CE as a philosophical method. The key issue here is to reflect more carefully on different varieties of control. It will emerge that although semantic externalism is incompatible with what I will call ‘individual immediate control’, it is well compatible with ‘collective long-range control’. Drawing on analogies to other normative domains such as the ethics of climate change, I argue that collective long-range control is enough to get CE off the ground.[[2]](#footnote-2)

Here is how I will proceed: In section 2, I will outline the externalist challenge to CE in some more detail. I will show how this challenge can be reconstructed as an argument with three premises, all of which will be discussed and critically assessed in the following sections. Section 3 discusses Herman Cappelen’s reasons for endorsing two of the premises but rejecting the third, which is the claim that CE requires meaning control (Cappelen 2018). Section 4 aims to show that Cappelen’s case against this premise is unconvincing and that there are independent reasons that speak in its favor. Sections 5, 6 and 7 provide an extended discussion of premise 2, i.e., the claim that semantic externalism precludes meaning control. Section 5 develops the notion of ‘collective long-range control’. Section 6 argues that semantic externalism can account for reference change. Finally, section 7 argues that the levers of reference change outlined in section 6 are within our collective long-range control and points to some plausible generalizations of this view.

**2 The Challenge**

*Semantic* theories assign semantic contents to expressions of a language. Their goal is to answer questions like ‘What is the meaning of a given expression *e* by a speaker *S*?’. *Metasemantic* theories state the facts in virtue of which expressions by a speaker have the semantic contents they have. Their goal is to answer questions like ‘Why is it the case that a given expression *e* by a speaker *S* has meaning *m*?’. The perhaps most important divide among metasemantic theories is that between internalism and externalism. The standard characterization of this divide goes as follows (cf. Wikforss 2008):

**Semantic Internalism (SI):** Facts about the meaning of a linguistic expression *e* of a speaker *S* are entirely grounded in facts about the mental states of *S*.[[3]](#footnote-3)

**Semantic Externalism (SE):** Facts about the meaning of a linguistic expression *e* of a speaker *S* are not entirely grounded in facts about the mental states of *S*; at least some of them are partly grounded in facts external to the mind of *S*.

Now consider two skeptical conjectures about the interrelation between SE and CE.

**Conjecture 1:** If (a certain branch of) SE is true, then the means that philosophers who engage in CE concerning some lexical item *t* have at their disposal are ineffective in changing *t*’s meaning.

*Rationale:* What can philosophers who engage in CE about *t* do to accomplish their goal? They can write papers giving reasons why it would be better if *t* had meaning *m*\* instead of *m*, they can talk about it to their colleagues, give presentations at conferences, etc. Even if their papers and talks are maximally convincing, the best they could ever achieve is to make other philosophers believe that it would be better if *t* had meaning *m*\* instead of *m* and to make their communicative intentions change accordingly. If internalism is true, this activism pays off. Having all members of a group intending to use *t* with meaning *m*\* (and eventually also believing that *t* means *m*\*) *makes it the case* that *t* means *m*\*. But if certain branches of SE are true, this activism is just a big waste of time.[[4]](#footnote-4) If the meaning of *t* is grounded in other things than one’s attitudes about *t*, then *t* may continue to have meaning *m* despite even large-scale changes of attitudes.

**Conjecture 2:** If (a certain branch of) SE is true, then changing the meaning of lexical items is either impossible or at least very hard to achieve.

*Rationale*: According to what has become known as the causal theory of reference, some of the facts which ground reference are *facts about the past.* On this theory, proper names have their reference determined by initial acts of dubbing, or ‘baptism’ (Kripke 1980, 96). Once the act of dubbing is completed, reference is established. This implies that at a given time *g* the meaning of a term *t* is determined by an event that took place at *g*-1. But if this is true, then how can we change *t*’s current meaning? Wouldn’t this require us to be able to change the past? Putnam and Kripke have also extended this theory to natural kind terms, claiming that once the reference of terms like ‘water’ or ‘gold’ is established, their meanings are constrained by what happens to be the microphysical structure of water and gold. If one were to change their meanings, it seems that one would therefore have to change either facts about the past, or necessary facts about gold and water. Since both kinds of changes are clearly beyond our control or even metaphysically impossible, it may seem that on certain prominent versions of SE, CE is actually infeasible.

Let us now turn this line of thought into a proper argument:

**Externalist Challenge (EC)**

1. SE is true about many terms of our language, and in particular about those terms typically in the focus of practitioners of CE.
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.

Some of the ingredients of EC are intentionally left vague in its present formulation. For example, what exactly are the terms typically in the focus of conceptual engineers? What does it take for something to be ‘within our control’? What does it mean for a method to be ‘not applicable’ to a case? The first two of these questions will be tackled as we go along. The third, however, points to an intentional blank space in the argument. The idea is that you can fill in for ‘not applicable’ in (3) and (4) whichever kind of negative property you deem adequate, such as ‘impossible’, ‘undoable’, ‘not something we should do’, ‘not a good philosophical method’ and potentially many more. The general claim that EC tries to establish is that SE poses some kind of serious problem for the method of CE, thus allowing for different ways to spell out exactly what this problem is.

Using the reconstruction of EC above, we can distinguish between three different strategies to circumvent it. One strategy is to reject premise (3), claiming that a lack of control does not stand in the way of CE. This strategy has recently been endorsed by Cappelen (2018) and it will be outlined and finally rejected in sections 3 and 4. Another strategy amounts to a rejection of (2), arguing that SE does in fact allow for a significant degree of meaning control. In my view, this is the most promising strategy and it will be pursued in sections 5-7.

Cappelen’s and my envisaged solutions to EC have one thing in common: they both assume the correctness of SE, that is, they both leave premise (1) of EC untouched. Of course, it is also possible simply to reject SE. If SE is false, then there is no externalist challenge. I will not pursue this option here. Instead, I will take it for granted that SE is influential enough to make its compatibility with CE a live and important methodological issue.

There is yet another way to reject (1) that should be mentioned. Classic Putnam-Kripke-style externalism is typically introduced to hold for proper names and natural kind terms such as ‘water’ or ‘gold’. But why should conceptual engineers care for the metasemantics of names or natural kind terms? Surely, they never meant to change *those*! Indeed, the cases of CE that are being discussed in the literature – ‘truth’ (Scharp 2007; Scharp 2013), ‘knowledge’ (Fassio and McKenna 2015), ‘woman’ (Haslanger 2000; Diaz-Leon 2016), ‘race’ (Haslanger 2000), ‘belief’ (Clark and Chalmers 1998) – do not seem to be on a par with ‘gold’ or ‘water’, and they aren’t proper names either. So, it seems that it is possible to grant traditional versions of SE without having to face EC after all.

This refutation of EC would be too quick, however. First off, for some of the entries on the above list, it is questionable whether they are natural kind terms. For example, Kornblith (2002) argues that knowledge is a natural kind, and insofar as ‘knowledge’ successfully picks it out, ‘knowledge’ might in fact be a natural kind term.[[5]](#footnote-5) Similarly, Clark and Chalmers (1998) suggest that there is a natural kind in the vicinity of ‘belief’. So chances are that at least some of the recently made engineering proposals do in fact involve natural kind terms. But even if they don’t, there is no systematic reason to restrict the method of CE in this way. After all, it might for some specific reasons be better to change the reference of a word from one natural kind to another, or from one natural kind to a non-natural kind.

Second, and perhaps more importantly, some authors have suggested that Putnam-Kripke-style externalism can be extended to other semantic types than names and natural kind terms. In fact, in *The Meaning of ‘Meaning’*, Putnam himself claims that it applies “to the great majority of all nouns, and to other parts of speech as well” (Putnam 1975, 242). The examples he gives include ‘pencils’, ‘bottles’ and ‘tables’. With respect to all of them, it might turn out that they have a hidden nature, i.e., essential properties of which we are unaware. If this is the case, then we could run Twin-Earth-style arguments for these artefact terms as well (Putnam 1975, 242–3).[[6]](#footnote-6) In a similar vein, Haslanger (2006) suggests that “externalist insights should be applied to our thought and language about the social as well as the natural” (395), and therefore generalizes *natural kind externalism* to *objective type externalism* (398). Proposals along these lines currently gain a lot of traction. For these reasons, restricting externalist insights to proper names and natural kind terms in order to safeguard CE from its implications does not seem like a fruitful theoretical choice.

**3 Cappelen’s Solution**

The second strategy to deal with EC that I want to consider here is to argue that even if we do in fact lack control over meanings, this does not stand in the way of pursuing CE as a philosophical method. Cappelen (2018) implements this strategy along the following lines. First, he presupposes (1), i.e., the truth of SE; second, he argues that the truth of SE implies (2), i.e., that we lack meaning control; third, he argues that a lack of meaning control does not stand in the way of pursuing CE, thereby rejecting (3) and thus the unwanted conclusion (4). In this section, I will sketch his case for (1) and (2) and an element of his view that implies the falsity of (3). In the next section, I will take a closer look at his arguments against (3) and argue that they fail. The conclusion of these two sections will be that, given Cappelen’s case for (1) and (2), there is no room for rejecting (3) and thus no way to avoid (4).

Here is how Cappelen summarizes his own metasemantic views[[7]](#footnote-7):

(a) *Large-scale externalism*: Intensions and extensions are partly determined by the external environment that speakers happen to be in. The relevant aspects of the external environment include *experts in the community* (Burge), *history of use* going back to the introduction of a term (Kripke), *patterns of use over time* (Kripke, Williamson), and *what the world happens to be like* (independently of what the speakers believe the world is like) (Putnam). Intensions and extensions don’t supervene only on the mental states of the speakers. (Cappelen 2018, 63)

(b) *Possibility of error:* Most or even all speakers of a language can *believe* that a predicate F applies to an object, o, but be wrong. They can all *want* o to be in the extension of F, but wanting o to be F doesn’t make it so. They can all be *disposed* to apply F to o even though o isn’t F. Believing and wanting words to mean something doesn’t make it so. In general, our attitudes towards the meaning of a term have little effect on its actual meaning. (Cappelen 2018, 63)

The picture endorsed by Cappelen is a stark form of SE, in three respects. First, it entails what we might call ‘global externalism’ – the idea that SE is true of all, or at least most, semantic types. Second, Cappelen’s version of SE is incompatible even with what is sometimes called ‘mentalist-externalism’, i.e., the idea that speaker meanings are grounded in the mental states of the speaker plus other participants of the speaker’s linguistic community. Cappelen’s position is incompatible with mentalism, for he deems the following scenario possible: all speakers of a language *L* *believe* that an expression *e* has a certain meaning, they all *want e* to have this meaning and they are all *disposed* to use *e* in accordance with this meaning, but nonetheless *e* means something different. Third, several passages suggest that when it comes to weighing the contributions to meanings made by mental states against the contributions made by non-mental factors, Cappelen believes that the former only play a minor role. Thus, he writes that “what people say, think, propose, wish and debate about meanings have *very little influence* on what words mean”, and where mental states do have an effect on meaning, this happens in “*unpredictable ways*” (Cappelen 2018, 59; my italics).

Against this backdrop, Cappelen makes two claims about meaning change: First, that SE generally allows for the possibility of meaning change; but, second, that the underlying mechanisms are to a large extent inscrutable and outside of our control. Cappelen cites discussions about reference change in the externalist literature and notes that there is nothing about SE which forbids reference shifts (Cappelen 2018, 65). However, he points out that there is no agreed-upon theory about how this happens, and diagnoses that this is no mere coincidence but symptomatic of a deeper issue: the relevant mechanisms are simply too heterogenous, too complicated and overall too messy to be captured by a theory. Just as there is no algorithm or recipe to extract meaning from use, he argues, so is there no algorithm or recipe for how to generate reference change (Cappelen 2018, 67).

Based on these metasemantic assumptions, Cappelen draws a few important consequences for CE, summarized in the following passage:

**Inscrutable – Lack of Control – Will Keep Trying:** The processes involved in conceptual engineering are for the most part inscrutable, and we lack control of them, but nonetheless we will and should keep trying. (Cappelen 2018, 72)

A bit more untangled, this view has three basic elements:

Inscrutable (epistemological):In most cases, the detailed mechanisms that underpin particular instances of CE are too complex, messy, non-systematic, amorphous, and unstable for us to fully grasp or understand.

Lack of Control (metaphysical): The process of CE is governed by factors that are not within our *control:* no one of us and no sub-group of us has any significant degree of control over how meaning change happens. Even if we could overcome our epistemic limitations – and know all about the relevant factors for a particular case – what we would have knowledge of would be something we had little control over.

Keep Trying (psychological?):Despite Inscrutable and Lack of Control, we *will* and *should* keep trying to engage in CE and given the kinds of creatures we are, maybe we *must* keep trying. (Cappelen 2018, 73)[[8]](#footnote-8)

In the next section, I will argue that Lack of Control and Keep Trying are *prima facie* incompatible. Furthermore, I will show that nothing Cappelen says in order to re-establish their compatibility is convincing. Linking this discussion to EC above, I read Lack of Control as Cappelen’s endorsement of premises (2) and (1), while Keep Trying implies the rejection of (3).

**4 Against Cappelen’s Solution**

The first thing to note about Keep Trying is that it contains three separate claims:

Will Keep Trying: We *will* keep trying to engage in CE.

Must Keep Trying: Given the kind of creatures we are, we *must* keep trying to engage in CE.

Should Keep Trying: We *should* keep trying to engage in CE.

Will Keep Trying and Must Keep Trying are empirical claims about human creatures. To my mind, both of them sound problematic. Saying that we will as a matter of fact keep trying to engage in CE seems like a rather speculative claim, to say the least (and nowhere in his book does Cappelen present any argument for it); but claiming that human beings are such that they *must* keep trying to engage in CE is not only speculative, but highly implausible. Surely, there are actual or at least possible human societies in which people neither care about CE nor even know what it is. But apart from being implausible, Will Keep Trying and Must Keep Trying are also irrelevant in the context of EC, for neither of them says anything about the viability of CE. Even if both were true, we would still not know whether CE is an applicable and valuable method of doing philosophy. This is why I will only focus on Should Keep Trying in the following.

The tension between Lack of Control on the one hand, and Should Keep Trying on the other seems fairly obvious. Lack of Control strongly suggests that changing the meaning of a term is something we cannot do – it implies that meanings are determined by factors we have no control over. So even if, pace Inscrutable, we did know how to change meanings, we still could not do it, at least not intentionally.

Now consider how this observation can be used in an argument against the idea that we should keep trying to engage in CE:

1. We should keep trying to engage in CE only if there is a realistic chance of success.
2. If we do not possess meaning control, then there is no realistic chance of success when engaging in CE.
3. We do not possess meaning control.
4. Therefore: there is no realistic chance of success when engaging in CE.
5. Therefore: it is not the case that we should keep trying to engage in CE.

I anticipate that (i) is the premise that Cappelen will disagree with. One apparently plausible way of doing this is by arguing that it rests on an equivocation of what it means to engage in CE: on the one hand, one does so simply *by evaluating concepts and suggesting improvements*; on the other hand, one does so *by actually changing the meanings* of terms/concepts. For each of these activities, it might be right that one ought to engage in them only if there are realistic chances of success. But this doesn’t entail that one ought to engage in the former only if there are realistic chances of succeeding in the latter.

In response, let me stress that conceptual engineers typically want their proposals to be implemented. Conceptual engineers think that our actual or traditional way of talking/thinking about a certain subject matter or of framing a philosophical or scientific question is less than ideal: either because it has bad epistemic or moral consequences, or because it gives rise to paradoxes, problematic cases of vagueness, etc. They seek to solve these problems by revising the concepts we use to phrase the subject matter in question. This project is successful only if the conceptual revisions are actually implemented. This is perhaps most vivid with respect to proponents of what have become known as ‘ameliorative projects’. What is distinctive of these projects is that they regard conceptual revision at least partly as a means to bring about social change. Conceptual engineers of this kind are both theorists and social activists at the same time: they propose new meanings of certain concepts in order to change the world in certain respects. And here it is obvious that their proposals crucially depend on the practical possibility of implementing their suggested revisions.[[9]](#footnote-9)

This being said, let us look at what Cappelen, who anticipates this kind of worry, says in his defense:

Some readers of early drafts of this manuscript found it weird that I present it as a defense of the importance of conceptual engineering, but at the same time defend theses like ‘Lack of Control—Inscrutable—Will Keep Trying’. They asked whether this book is in effect a debunking project, disguised misleadingly as a defense and whether this book is in effect an extended argument that conceptual engineering is impossible. I understand this reaction, but it’s a mistake for reasons I’ve already mentioned. As I emphasized above, that the processes underlying change are inscrutable and uncontrollable is a familiar point from other normative domains. Theorists are comfortable reflecting on and proposing theories of justice, for example, without having a recipe for how they can be implemented. More generally, we can make judgments about what ought to be the case without knowing how to make the world that way (or even having a plausible strategy in mind). Those are just different projects: figuring out what ought to be the case and making the world that way. (Cappelen 2018, 83-4)

I think that Cappelen’s claim that the value of normative theories does generally not depend on knowledge about how to implement them can hardly be resisted. However, this claim is insufficient to establish what Cappelen set out to establish, namely that CE is a worthwhile activity despite the fact that it is inscrutable and out of control. There is a difference between proposing theories whilst *not having a recipe for how they can be implemented* on the one hand, and proposing theories whilst *knowing that they cannot be implemented* on the other. The first is compatible with the idea that theories provide blueprints for action. For example, from reflections on what an ideal structure of society would look like, we can, at least in principle, derive guidelines for concrete changes of the status quo. The theory is valuable because it gives us orientation. However, if a theory comes with the knowledge that, for principled reasons, it cannot be implemented, or at least not to a satisfying degree, then it cannot be regarded as a blueprint for action and it cannot provide us with orientation either. But what, then, is the value of the theory in the first place?

Cappelen seems to assume that in claiming that CE is free of any feasibility constraints, he is in the good company of political philosophers who have been engaging in ideal theory all along.[[10]](#footnote-10) This, however, is only partly true. It is true that in Rawls’ *A Theory of Justice* (Rawls 1999a) idealizations play a major role in developing his theory of justice. He there introduces the distinction between ideal and non-ideal theory, claiming that the ideal theory of justice gives an account of what he elsewhere calls a ‘realistic utopia’ (Rawls 1999b), i.e., the best we can realistically hope for, whereas the non-ideal theory asks how this long-term goal can be achieved in the realm of actual politics. In the words of Rawls, non-ideal theory thus “looks for courses of action that are morally permissible and politically possible as well as likely to be effective” (Rawls 1999b, 89). To make the analogy to Rawls work, Cappelen would have to supplement his ideal theory with a non-ideal theory which states realistic courses of action that can be undertaken in order to bring about conceptual change. But he doesn’t. Moreover, Rawls makes it clear that even the principles of ideal theory must be adjusted to the limits of political possibility. The objective is to “probe[] the limits of practicable political possibility” (Rawls 2001, 4) by making only realistic assumptions in our ideal theory. This is necessary to avoid “idle utopianism” (Simmons 2010, 8). It thus seems that Cappelen’s dismissal of feasibility constraints cannot be based on parallel ways of reasoning in political philosophy either.[[11]](#footnote-11)

Putting Cappelen’s arguments to one side, I think it is worth highlighting the significance of what we are currently discussing. For CE to play an important role in philosophical theorizing – as both Cappelen and I believe[[12]](#footnote-12) – we need to show more than that reference shifts are possible. We need to show that at least some minimal form of optimism about our capacity to influence what our terms or concepts mean is warranted. If no such optimism can be had, then CE is not properly conceived of as a method of doing philosophy. For something to be a method of an academic discipline it must be such that one can intentionally apply it to a given case and be rational in expecting that the application has a chance of being successful. If Cappelen is right, then this is not the case for CE, because Lack of Control entails that we have no reason to expect that our efforts will be successful – or even have any effect at all.

**5 Varieties of Control**

If what I said in the above is true, refuting EC on the grounds of a rejection of premise (3) is not successful. In this section and the following ones, I will pursue a more promising option, namely to consider more carefully whether SE actually entails a lack of control, i.e., whether premise (2) of EC is correct. I will argue, first, that even externalist theories like the causal theory of reference leave us with what I will call ‘collective long-range control’ over word meanings and, second, that this is sufficient to make a case against (2). What I will do in this section is to develop the notion of collective long-range control; in the next sections, I will argue that causal theories of reference typically allow for this kind of control over word meanings.

Let us start with a fresh look at the notion of ‘being in control’. This notion surfaced many times throughout this paper. But what does it take to be in control of some state of affairs? Surprisingly, Cappelen, who argues at length that we do not possess control over word meanings, never stops to identify more clearly what this even means. What is more, he even explicitly denies that this notion stands in need of clarification: “’Significant’, ‘impact’ and ‘control’ are all perfectly respectable English words, I trust readers to understand them. There are no doubt senses of ‘control’ different to that which I intend but those are not the ones that I use” (Cappelen 2018, 158). I am not convinced that this is true. Sure, we do have some understanding of what it means to be in control of something. But our ability to grasp such expressions doesn’t give us enough to handle the present, much more abstract and much less ordinary case. In fact, even reflecting on rather mundane cases quickly reveals that control comes in different varieties.

Consider the following distinction first: There is a clear sense in which I am in control of which letters I am typing right now. If I form the intention to type the letter ‘r’, nothing hinders me from doing it. This is not to say that nothing could possibly interfere – my computer might shut down or simply refuse to take in the letters I am typing. But nonetheless, given my intention, chances of success are sufficiently high to say that I am in control. Things look a bit different when it comes to body weight. There is no single uninterrupted action I can take to significantly reduce or increase my body weight (at least not in a healthy way!). Whatever I can do to affect my body weight will take a significant amount of time. Nonetheless, it seems right to say that I do have a certain amount of control over my actual body weight.

William P. Alston (1988; 2005) once suggested to capture this distinction with definitions along the following lines:

**Immediate Control**: *S* has immediate control over some condition *c* iff there is a single uninterrupted action *S* can take that will, with a sufficient degree of likelihood, have *c* as a consequence.[[13]](#footnote-13)

Examples: raising my arm, turning on the light switch.

**Long-range Control**: *S* has long-range control over some condition *c* iff there is a series of potentially interrupted actions, to be performed over a potentially significant period of time, that *S* can take and that will, with a sufficient degree of likelihood, have *c* as a consequence.

Examples: reducing one's body weight/cholesterol concentration/blood pressure, changing one’s behavioral dispositions.[[14]](#footnote-14)

Notice that immediate control implies long-range control, but not *vice versa*. This is because the conditions on immediate control are stricter than those on long-range control. If there is a single uninterrupted action which results in c, then there is also a series of potentially interrupted actions which results in c.

There is a further distinction I would like to introduce. Consider the following scenario: An election takes place in your country and each person with the right to vote gets one vote. You as an individual neither have immediate control nor long-range control over who wins the election. Sure, you do have *some* degree of influence – otherwise, why would you go voting in the first place? –, but this degree is insufficient to ascribe control to you. Nonetheless, it wouldn’t be right to say that the outcome of an election is completely beyond anyone’s control, for even though there is no individual who is in control, the voting collective as a whole *is*. To capture this difference, we must distinguish between *individual* and *collective* control.

Both individual and collective control can be immediate and long-range. The election case is an example of immediate collective control: to make it the case that a certain party wins the election, all that is necessary is that sufficiently many people perform a single uninterrupted action, namely to put a cross at the right place on the voting sheet. Reaching certain climate goals, on the other hand, may be said to be something that we, the human species, have collective long-range control over. For instance, significantly lowering the carbon dioxide and methane outputs is nothing that an individual can achieve; neither is it something that a collective can achieve by performing a single uninterrupted action. To reach climate goals, we have to take a course of action and perform it over a significant period of time. The same is true of many other global long term goals, such as ending world poverty, putting an end to gender inequality, providing clean water and sanitation for everyone, etc. All these goals describe states of affairs that we at most have collective long-range control over. Putting the two distinctions together thus gives us two more varieties of control:

**Collective Immediate Control**: A group *G* has immediate control over some condition *c* iff there is a (set of) single uninterrupted action(s) that sufficiently many members of *G* can take that will, with a sufficient degree of likelihood, have *c* as a consequence.

**Collective Long-range Control**: A group *G* has long-range control over some condition *c* iff there is a (set of) series of potentially interrupted actions, to be performed over a potentially significant period of time, that sufficiently many members of *G* can take that will, with a sufficient degree of likelihood, have *c* as a consequence.

Again, individual control of a certain kind implies collective control of that same kind, but not *vice versa*. If there is a single agent who is in control of c, then there is also a group who is – namely the group which consists of this agent only.Notice further that, as cases like the fight against climate change or gender inequality show, the fact that we merely possess collective long-range control over some condition *c* does not generally undermine the rationality of conducting a normative project about *c*.

With this taxonomy in mind, we can now take a closer look at how SE can handle reference change (section 6) and whether the processes involved therein can be said to be within our control (section 7).

**6 Semantic Externalism and Reference Change**

A difficulty in assessing the kind of meaning control that SE allows for is that there are many different externalist theories, all of which may differ with respect to the matter at issue here. Within the limits of this paper, it will not be possible to take a look at *all* of them. What I will do instead is to focus on the most classic and generally most popular themes of the externalist literature – natural kind externalism as championed in Putnam (1975) and the causal theory of reference as introduced in Kripke (1980) and further developed in Evans (1973), Devitt (1981) and Devitt and Sterelny (1999). Although assessing these versions of SE will certainly not justify the claim that *all* versions of SE allow for meaning control, I submit that it brings us a great deal forward. In this section, I will try to establish two major theses: (i) that cases of reference change are systematically plausible, and (ii) that causal theories of reference have the resources to account for them. Finally, in the next section, I will argue that the mechanisms used in these accounts are within our collective long-range control.

Let us start with two intuitive thought experiments, both of which are versions of Hilary Putnam’s famous Twin Earth case (Putnam 1973; 1975):

*Young-Mary on Twin Earth*

We are back in 1750 when people had very little knowledge of chemistry. Elsewhere in the galaxy there is a planet exactly like Earth in all respects except that the watery stuff that fills the oceans, rivers and lakes is not composed of H2O but has a very long and complicated formula abbreviated as ‘XYZ’. Call this planet ‘Twin Earth’. Now suppose that some Earthians build a spaceship and travel to Twin Earth. After landing on Twin Earth, Mary, who is one of the crew members, quickly recognizes the planet’s resemblance with Earth. When the crew reaches the shore of what appears to be a great lake, she says to one of her crew members: ‘Unbelievable! There is even water on this planet!’.

*Old-Mary on Twin Earth*

Now suppose that Mary and some of the other crew members decide to permanently settle on Twin Earth. As time progresses, more and more Earthians do the same. All of them do all the stuff with twin-water that normal Earthians do with water: they drink it, shower with it, swim in it, etc. None of them is aware of any difference between water and twin-water. They all keep calling twin-water ‘water’. Towards the end of her long life that she has spent mostly on Twin Earth, Mary goes back to the same shore she and her crew found right after landing on Twin Earth, and mumbles to herself: ‘There is so much water here. I still remember how excited I was when I first saw it.’

Here is what I take to be the intuitive verdicts about these cases: In young-Mary’s mouth, ‘water’ refers to water, not to twin-water. Given that there is no water on Twin Earth, her utterance is therefore false. In old-Mary’s mouth, however, ‘water’ refers to twin-water. Given that there is twin-water in the lake before her, her utterance is therefore true. If some further vindication of these verdicts should be needed, think of it this way: there is a clear sense in which young-Mary’s use of ‘water’ is discontinuous with how she and her fellow travelers have used the term before; old-Mary’s use of ‘water’, on the other hand, is perfectly continuous with how she and her fellows on Twin Earth have used the term for the better parts of their lives. If one agrees with these verdicts (or if one at least concedes that there are possible pairs of examples of this kind), then one must also agree that the reference of ‘water’ in Mary’s mouth has changed at some point in her life.[[15]](#footnote-15)

Now let us see whether causal theories of reference can accommodate this change of reference and what they have to say about the mechanism behind it. I will begin with Kripke’s version of the causal theory and then turn to Evans’ and Devitt’s respective variations of it. Kripke’s version of the causal theory, as developed in *Naming and Necessity*, involves two crucial elements:

(1) *Baptism*. “An initial 'baptism' takes place. Here the object may be named by ostension, or the reference of the name may be fixed by a description.” (Kripke 1980, 96)

(2) *Causal chain*. “When the name is 'passed from link to link', the receiver of the name must, I think, intend when he learns it to use it with the same reference as the man from whom he heard it.” (Kripke 1980, 96)

Although this view is primarily introduced as an account of proper names, Kripke makes clear that it holds for natural kind terms as well (Kripke 1980, 139). This view nicely explains the intuitive verdict about *Young-Mary on Twin Earth*: We may suppose that ‘water’ was initially introduced to name water. As a matter of fact, water is composed of H2O. Furthermore, the chemical structure of natural kinds like water or gold is essential to their identity: something that is not composed of H2O just isn’t water (Kripke 1980, 124-134). Young-Mary’s use of ‘water’ constitutes a link in a causal chain that ultimately terminates in an event of baptism that ties ‘water’ to water, because it seems plausible that young-Mary intends to use ‘water’ with the same reference as the person from whom she acquired it. For this reason, ‘water’ in her mouth refers to water; and given that the stuff on Twin Earth is not composed of H2O, it isn’t water. Hence her utterance is false.

However, it seems that this view has difficulties in predicting the intuitive verdict about *Old-Mary on Twin Earth.* For Kripke, what matters for reference preservation is the person’s *intention* to use the term exactly as it was passed on to her. But this condition seems to be satisfied by both young- *and* old-Mary. Neither Mary, nor anybody else on Twin Earth ever discovered a difference between water and twin-water. Mary and her fellow Twin Earthians always had the intention to use ‘water’ exactly as it was transmitted to them (or so we may suppose). But then why does old-Mary refer to twin-water, whereas young-Mary refers to water?

At the very end of *Naming and Necessity,* Kripke acknowledges the possibility of reference change and briefly considers the explanation that “a present intention to refer to a given entity…overrides the original intention to preserve reference in the historical chain of transmission” (Kripke 1980, 163). Although this explanation might be on the right track, it lacks the detail that is required to account for cases like the above, for Kripke does not specify the conditions that have to be satisfied for such ‘overriding’ to take place. Indeed, Kripke himself admits that “the matter deserves extended discussion”, that it “requires more apparatus than [he has] developed here” and therefore suggests leaving the problem for further work (Kripke 1980, 163).

Kripke himself has not conducted such further work, but Gareth Evans and Michael Devitt have. On the account Evans ends up with, the reference of a proper name as used by *S* is “the source of causal origin of the body of information that *S* has associated with the name” (Evans 1973, 198). Evans thus shares Kripke’s idea that reference supervenes on the causal relation between the object, or person, and the speaker; but according to Evans, Kripke “has mislocated the causal relation; the important causal relation lies between that item’s states and doings and the speaker’s body of information – not between the item’s being dubbed with a name and the speaker’s contemporary use of it” (Evans 1973, 197). Similarly, on Devitt’s account, the reference of a proper name as used by *S* is the object that grounds those thoughts of *S* which dispose *S* to use the name (Devitt 1981, 131). Here, too, the important causal relation is the one between the object and a particular subset of *S*’ mental states.

Accounts like Evans’ and Devitt’s neatly explain the difference between young-Mary and old-Mary in the examples above. In Evans’ terminology: young-Mary refers to water because all, or by far the most, pieces of information she associates with ‘water’ have water as their causal source. This makes sense, given that, up to this point in time, she has lived most of her life on a planet that contains only water and that she has communicated only with people who have no connection whatsoever with twin-water. Old-Mary, on the other hand, spent most of her life on Twin Earth. By far the most of her interactions with watery stuff involved twin-water. The same is true of most of the utterances she overheard on Twin Earth: when her friends used the term ‘water’ to pass on information to her, these pieces of information also had twin-water as their causal source. This is why old-Mary refers to twin-water when using the term ‘water’.[[16]](#footnote-16)

Apart from giving us the correct verdicts about the two cases above, Evans’ and particularly Devitt’s account provide some illuminating details about the mechanisms of reference change. The key idea is that the properties of being the causal source of someone’s body of information associated with a term or of being the ground of one’s token-involving thoughts are *gradable* properties. It is thus possible that an object *o*1 has the property of being the causal source for the body of information that *S* associates with a term *t* to the degree *k* (1 ≥ *k* ≥ 0), and that *o*2 has this property to the degree 1-*k* (and similarly for Devitt’s account). This is plausible because there are lots of different (types of) influences on what happens to be the causal source of the body of information one associates with a term. For example, every time a person sees, feels or hears water, every time she reads about it or listens to others who talk about it, this will influence her causal source of ‘water’. If some of those influences come from objects that are not water, but, say, twin-water, then twin-water will enter the causal source of the person’s body of information she associates with ‘water’. Devitt describes these cases as cases of *partial reference*: a person whose *t*-involving thoughts are grounded in object *o*1 to the degree *k*, but to object *o*2 to the degree 1-*k*, will refer to *o*1 to the degree *k* and to *o*2 to the degree 1-*k* when using *t* (Devitt 1981, 147). On these views, reference change therefore turns out to be a gradual and complex process, to which lots of different events may contribute.

**7 Reference Change and Meaning Control**

The only remaining question is whether the process just described can reasonably be said to be within our control. I believe that the answer is ‘yes’ – at least in the sense of what I have earlier identified as *collective long-range control*. On Evans’ and Devitt’s views, what it takes to bring about reference change is to make it the case that a new entity becomes the *dominant* causal source or ground of a particular subset of one’s mental states – namely those that are associated with the term in question. In cases where the two objects are superficially identical, like water and twin-water, the internal properties of these mental states will remain the same; the only thing that changes is their causal source. In cases where the two objects are not superficially identical, this process will also be accompanied by a change of internal mental states. What is crucial is that the process just described is one over which we possess collective long-range control*.* Recall that collective long-range control over some condition *c* requires that there be a series of actions, to be performed over a potentially significant period of time, such that if sufficiently many members of a group perform it, it will likely have *c* as a consequence. This requirement is fulfilled with respect to changing what is the dominant causal source of the body of information we associate with a term. Here is, roughly, how this might look like: 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. Likewise, some of the information caused by the previous object or kind will be lost from the body of information. Thus, little by little, the term will shift from the old reference to the new one, until the new object or kind has gained overall dominance.[[17]](#footnote-17)

For illustration, consider a version of the Twin Earth case in which Mary and her fellow crew members start investigating the microphysical structure of twin-water and learn that it is not composed of H2O but of XYZ. Suppose they agree that, because of this fact, twin-water is not in the extension of ‘water’; but instead of introducing a new term for twin-water, they decide to work towards a reference change of ‘water’. Freshly landed on Twin Earth, by far the most of their experiences involved water, which is why the bulk of the information they associate with ‘water’ has water as their causal source. Nonetheless, they go on to apply ‘water’ to twin-water. Scientists on Twin Earth who write about water base their observations on twin-water; doctors use the word ‘water’ to tell their patients to drink more twin-water; parents tell their children not to swim in ‘deep waters’; etc. Simultaneously, the information people had associated with ‘water’ before they came to Twin Earth fade. In this manner, more and more of the information they associate with ‘water’ becomes causally grounded in twin-water. Eventually, ‘water’ will switch reference from water to twin-water.

For the purposes of my argument, it is not necessary to show that causal theories of reference give us collective long-range control *only*; it might well be that there are cases in which people possess one of the stronger forms of control as well. That being said, I submit that in the great majority of cases we should not expect more than collective long-range control. This is because, on causal theories of reference of the considered kind, bringing about a change of reference usually requires others to play along, as their language use will influence what our own utterances refer to. Suppose, for instance, that nobody would follow Mary in using ‘water’ for twin-water. This would make it close to impossible for Mary to make twin-water the dominant causal source of even her own ‘water’ associations, much less of those of the speech community as a whole. So, it seems that in the case above, we should not expect any significant individual control. Immediate control seems equally unlikely. Suppose Mary and her fellows on Twin Earth came together and decided that from now on, ‘water’ means twin-water instead of water. Would that immediately make it so? According to causal theories in the spirit of Evans and Devitt, it would not – collectively forming the intention to use ‘water’ for twin-water will not immediately make twin-water the dominant causal source of people’s ‘water’ associations.

Before wrapping up, let me briefly discuss an objection one might have against this view. Above, I illustrated how a term can switch reference. It is a crucial feature of this account that the old reference and the new one do not overlap too much in the world of the speaker – otherwise, it would not be possible to influence reference by acts of grounding. For example, what explains the change of reference between young-Mary and old-Mary is the fact that Mary repeatedly has causal contact with a substance that is XYZ but not H2O. However, one might think that this feature of the example is not present in some of the conceptual reforms that have been advocated in the literature. For instance, Haslanger’s (2000) suggested understanding of ‘woman’ – roughly: *female human being who is oppressed on sex-marked grounds* – has more or less the same extension as our traditional concept of woman – roughly: *female human being*.[[18]](#footnote-18) The mechanism I described above may seem unfit to account for this case, for if the two concepts are co-extensional*,* then how can causal interaction with instances of the former concept lead to a change in meaning?

In response, I think that advocates of the causal theory of reference can make two (related) suggestions. First, on Evans’ version of the causal theory, one of the two relata of the relevant causal relation is the body of information that a speaker associates with the term in question, and it is of course possible that this body of information undergoes significant changes while its causal source of origin remains constant. In other words, Evans’ account allows us to change our *beliefs* about women and to preserve the *reference* of ‘woman’ at the same time. Although this does not amount to reference change, belief revision of this kind might often turn out to be what matters.

However, I submit that some conceptual engineers aim for neither extensional reference change nor mere belief revision, but for revision of intensions. So how can advocates of the causal theory of reference account for changes of intensions where the actual extension remains constant? To answer this question, it seems helpful to consider what has become known as the ‘qua problem’ for causal theories of reference. Devitt and Sterelny (1999, 91) give the following exposition:

The term is applied to the sample […] *qua* member of one particular natural kind. Any sample of a natural kind is likely to be a sample of many natural kinds; for example, the sample is not only an echidna, but also a monotreme, a mammal, a vertebrate, and so on. In virtue of what is the grounding in it *qua* member of one natural kind and not another? As a result of groundings, a term refers to all objects having the same underlying nature as the objects in the sample. But which underlying nature? The sample shares many.

Our current problem is related to the *qua* problem: Suppose every member of the kind *female human being* is also a member of the kind *female human being who is oppressed on sex-marked grounds.* Presumably, the former is a biological kind, while the latter is a social kind. Haslanger and her followers want to make it the case that ‘women’ applies to a sample of women *qua* member of the social kind, instead of *qua* member of the biological kind. To see how this can be done, we need to know which factors determine the relevant type of a sample; in other words, we need a solution to the *qua* problem.

Solving the *qua* problem goes beyond the scope of the present paper, so I will merely apply a suggestion that has been given in the literature. According to Devitt and Sterelny (1999), the indeterminacy posed by the *qua* problem is settled by *descriptive content*: “The relevant nature is the one that is, as a matter of fact, picked out by the descriptions associated with the term in the grounding” (92).[[19]](#footnote-19) The resulting theory is, of course, not a *pure* causal theory of reference, but a *hybrid* theory which involves both causal and descriptive elements. However, we have seen already that causal theories do good to include descriptive contents, as this helps them to account for reference change. So an advantage of the present proposal for solving the qua problem is that it only makes use of resources which are already available. If this, or something along these lines, is correct, then CE for terms with identical or near-identical extensions works by changing the descriptive content we associate with the term in question. For example, in order to make ‘woman’ apply to the social kind suggested by Haslanger, we have to make it the case that people associate descriptions with the term ‘woman’ that mark it as a social kind of the envisaged sort. Doing this may not be an easy task, of course, but it is something about which we, as a linguistic community, possess collective long-range control.

The resulting picture is one on which *use* and associated content turn out to be the decisive levers for CE. Whereas pure reference changes can be affected by changes of use directed at actual extensions, changes of intensions which go beyond changes in actual extension further require a change in associated content. Moreover, causal theories of the considered kind show that effecting a shift of reference typically involves larger collectives and requires continuous communicative activities. To those who thought that meaning change can be achieved by a single philosopher in her armchair, this result will surely seem disappointing. But when we consider that the conceptual engineer’s goal is not simply to change what she herself refers to by using a piece of language, but to bring about a change of reference in her linguistic community (or a subgroup thereof), then these consequences of SE seem quite reasonable. Causal theories of reference thus give us no reason to doubt the feasibility of CE.[[20]](#footnote-20)

**8 Conclusion**

Let me conclude by way of a summary. The externalist challenge to conceptual engineering raises doubts about the compatibility of the latter with standard versions of semantic externalism. Put roughly, the worry is that if certain branches of externalism are true, then we simply lack the kind of meaning control that conceptual engineering seems to presuppose.

In this paper, I have addressed a number of different responses to this challenge. According to the first, externalism is simply not true of the semantic types which conceptual engineers care about. For instance, causal theories of reference are sometimes endorsed only for proper names and natural kind terms. However, I have raised two worries for this kind of response: first, that prominent advocates of externalism like Putnam suggest that the scope of externalism stretches beyond natural kind terms, and second, that it is far from clear why conceptual engineers should, in principle, steer around natural kind terms. As a matter of fact, there seems to be some overlap between the list of concepts about which philosophers have suggested that they are natural kind terms, and the list of concepts about which philosophers have made engineering proposals (e.g. ‘knowledge’, ‘belief’, ‘woman’).

The second strategy that I have discussed is the one recently championed by Cappelen (2018). According to his *Inscrutable – Lack of Control – Will Keep* Trying thesis, a lack of meaning control does not stand in the way of practicing conceptual engineering. I have raised some serious difficulties for this view.

For one, while Cappelen is certainly right that proposing better meanings for concepts does not commit one to having a clear recipe for how these meanings can be implemented, it nevertheless seems that in order for CE to play an important role in philosophical theorizing, we need to make sure that meaning change is practically feasible. Moreover, appealing to the ideal vs. non-ideal distinction in political philosophy does not really help to dodge such feasibility considerations, for there is a great consensus in political philosophy that feasibility considerations are normatively relevant.

What I have argued for instead is to endorse a third strategy against the externalist challenge, namely to assess whether semantic externalism really does preclude meaning control. It was my goal to establish that this is not the case. My argument for this claim involved two steps: Firstly, I have distinguished between different varieties of control and identified what I call *collective long-range control*. Roughly, to possess this kind of control over a state of affairs *c*, it is sufficient that there be a series of (possibly interrupted) actions that a group of people can perform over a period of time which will likely have *c* as a consequence. Considerations about reaching certain climate goals or ending world poverty show that we do generally think that this kind of control is enough to motivate normative projects. Secondly, I have argued that causal theories of reference do allow for this kind of control over reference change. Drawing on Gareth Evans’ and Michael Devitt’s versions of the causal theory of reference, I have shown that changing the causal grounds of a speaker’s body of information associated with a term is a gradual process that can be achieved e.g. through communication and direct acquaintance. Lastly, I briefly discussed how causal theories of reference can account for conceptual engineering with constant references and I pointed to some plausible generalizations of the view defended here.

To my mind, the overall result of this paper supports a modest form of optimism about conceptual engineering as a philosophical method. For even though reference change on causal theories of reference turns out to require a collective long-term effort, it is nevertheless something that we, as a linguistic community, can bring about willingly. Those who thought that changing language was something that could be done by the philosopher alone from the armchair might not be happy to embrace this result – but in my view, we should be careful to avoid this kind of hubris anyway.

Acknowledgements Research presented in this paper was funded by Deutsche Forschungsgemeinschaft

(D73285Z) and conducted as part of the research project “Disagreement in Philosophy” at the University

of Cologne. It has further profited from critical discussions at the Second Cologne-Leuven Epistemology

Meeting 2016 in Leuven and at the European Epistemology Network Meeting 2016 in Paris. Thanks also go

to Thomas Grundmann, Jens Kipper, Joachim Horvath, Dominik Balg, Steffen Koch, Anja Wieben-James

and John Henry James for invaluable feedback.

Acknowledgements Research presented in this paper was funded by Deutsche Forschungsgemeinschaft

(D73285Z) and conducted as part of the research project “Disagreement in Philosophy” at the University

of Cologne. It has further profited from critical discussions at the Second Cologne-Leuven Epistemology

Meeting 2016 in Leuven and at the European Epistemology Network Meeting 2016 in Paris. Thanks also go

to Thomas Grundmann, Jens Kipper, Joachim Horvath, Dominik Balg, Steffen Koch, Anja Wieben-James

and John Henry James for invaluable feedback

**Acknowledgements** My work on this paper was supported by a grant of Deutsche Forschungsgemeinschaft for the Emmy Noether Independent Junior Research Group *Experimental Philosophy and the Method of Cases: Theoretical Foundations, Responses, and Alternatives* (EXTRA), project number 391304769. Earlier versions of this material were presented at the *Externalism and Conceptual Change* workshop at University of St. Andrews (2018) and at the *Foundations of Conceptual Engineering* conference at New York University (2018). I would like to thank the audiences of these conferences for their helpful feedback. Special thanks goes to Joachim Horvath and Thomas Grundmann, who provided invaluable comments and suggestions at various stages from the first draft to the final version. Lastly, I am indebted to two anonymous referees of this journal, whose quick and constructive feedback helped to further improve this paper.

**References**

Alston, W.P. (1988). The Deontological Conception of Epistemic Justification. *Philosophical Perspectives*, 2: 257–299.

—— (2005). *Beyond "justification"*: *Dimensions of epistemic evaluation*. Ithaca: Cornell University Press.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy*, 4/1: 73–121.

—— (1986a). Individualism and Psychology. *The Philosophical Review*, 95/1: 3–45.

—— (1986b). Intellectual Norms and Foundations of Mind *The Journal of Philosophy*, 83/12: 697–720.

—— (1990). Wherein is Language Social?. In A. George (Ed.), *Reflections on Chomsky*, 175–91. Oxford: Blackwell.

Burgess, A. and Plunkett, D. (2013a). Conceptual Ethics I. *Philosophy Compass*, 8/12: 1091–1101.

—— (2013b). Conceptual Ethics II. *Philosophy Compass*, 8/12: 1102–1110.

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

Chomsky, N. (2000). New Horizons in the Study of Language and Mind. Cambridge: Cambridge University Press.

Clark, A. and Chalmers, D. (1998). The Extended Mind. *Analysis*, 58/1: 7–19.

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

Devitt, M. and Sterelny, K. (1999). *Language and reality*: *An introduction to the philosophy of language* 2. ed. Oxford: Blackwell.

Diaz-Leon, E. (2016). Woman as a Politically Significant Term. A Solution to the Puzzle. *Hypatia*, 31/2: 245–258.

Dorr, C. and Hawthorne, J. (2014). Naturalness. In D.W. Zimmerman and K. Bennett (Eds.), *Oxford studies in metaphysics*. *Volume 8*, 3–77. Oxford: Oxford University Press.

Eklund, M. (2015). Intuitions, Conceptual Engineering, and Conceptual Fixed Points. In C. Daly (Ed.), *The Palgrave handbook of philosophical methods*, 362–85. Basingstoke: Palgrave Macmillan.

Evans, G. (1973). The Causal Theory of Names. *Aristotelian Society Supplementary Volume*, 47/1: 187–225.

Farkas, K. (2008). *The subject's point of view*. Oxford: Oxford University Press.

Fassio, D. and McKenna, R. (2015). Revisionary Epistemology. *Inquiry*, 58/7-8: 755–779.

Field, H. (1973). Theory Change and The Indeterminacy of Reference. *The Journal of Philosophy*, 70/14: 462.

Haslanger, S. (2000). Gender and Race. (What) Are They? (What) Do We Want Them To Be?’. *Nous*, 34/1: 31–55.

—— (2006). What Good are Our Intuitions?. *Aristotelian Society Supplementary Volume*, 80/1: 89–118.

—— (2016). Theorizing with a Purpose. The Many Kinds of Sex. In C. Kendig (Ed.). *Natural kinds and classification in scientific practice*, 129–45. Abingdon Oxon, La Vergne: Routledge Taylor & Francis Group.

Horvath, J. (2016). Conceptual analysis and natural kinds: the case of knowledge. *Synthese* 193 (1): 167-184.

Kornblith, H. (2002). *Knowledge and its place in nature.* Oxford: Clarendon Press.

Kripke, S. A. (1980). *Naming and Necessity*. Oxford: Basil Blackwell Ltd.

LaPorte, J. (2004). *Natural kinds and conceptual change*. Cambridge: Cambridge University Press.

Lewis, D. K. (1999). New work for a theory of universals. In D.K. Lewis (Ed.), *Papers in metaphysics and epistemology*, 8–55. Cambridge: Cambridge University Press.

—— (Ed.), (1999) *Papers in metaphysics and epistemology.* Cambridge: Cambridge University Press.

Ludlow, P. (2014). *Living words*: *Meaning underdetermination and the dynamic lexicon.* Oxford: Oxford University Press.

Machery, E. (2017). *Philosophy within its proper bounds.* Oxford: Oxford University Press.

Mendola, J. (2008). *Anti-externalism.* Oxford: Oxford University Press.

Mikkola, Mari (2009). Gender Concepts and Intuitions. In *Canadian Journal of Philosophy* 39 (4): 559–584.

Mills, C.W. (2005). 'Ideal Theory' as ideology. *Hypatia: A Journal of Feminist Philosophy*, 20/3: 165–184.

Putnam, H. (1973). Meaning and Reference. *The Journal of Philosophy*, 70/19: 699.

—— (1975). The Meaning of 'Meaning'. In H. Putnam (Ed.), *Philosophical Paper Vol. 2. Mind, language and reality*, 215–71. Cambridge: Cambridge University Press.

Rawls, J. (1999a). *A theory of justice.* Cambridge, Mass.: Belknap Press of Harvard University Press.

—— (1999b). *The law of peoples*: *With "The idea of public reason revisited".* Cambridge Mass.: Harvard University Press.

—— (2001). *Justice as fairness*: *A restatement*. Cambridge Mass.: Belknap Press.

Reimer, M. and Michaelson, E. (2017). Reference. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Spring 2017.

Richard, M. (ms). The A- project and the B-project. Available here: <https://markrichardphilosophy.wordpress.com/work-in-progress/>

Scharp, K. (2007). Replacing Truth. *Inquiry*, 50/6: 606–621.

—— (2013). *Replacing Truth.* Oxford: Oxford University Press.

Schwarz, W. (2013). Against Magnetism. *Australasian Journal of Philosophy*, 92/1: 17–36.

Sen, A. (2006). What Do We Want from a Theory of Justice?. *Journal of Philosophy*, 103/5: 215–238.

Simmons, A. J. (2010). Ideal and Nonideal Theory. *Philosophy & Public Affairs*, 38/1: 5–36.

Speaks, J. (2017). Theories of Meaning. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Spring 2017.

Thomasson, A. L. (2007). *Ordinary Objects*. Oxford: Oxford University Press.

Valentini, L. (2012). Ideal vs. Non-ideal Theory. A Conceptual Map. *Philosophy Compass*, 7/9: 654–664.

Wikforss, A. (2008). Semantic Externalism and Psychological Externalism. *Philosophy Compass*, 3/1: 158–181.

1. See Burgess and Plunkett (2013a) for a similar outline of this challenge. [↑](#footnote-ref-1)
2. Some philosophers argue that externalism also applies to the level of content (cf. Burge 1979, 1986a). If this is true, a version of the externalist challenge as well as my way of responding to it might also apply to mental content. Although I am open to the idea that this is indeed so, I will not pursue this line of thought in this paper. [↑](#footnote-ref-2)
3. For current purposes, nothing hangs on whether this distinction is formulated in terms of grounding or supervenience. Everything I say in this paper could be rephrased in terms of supervenience. [↑](#footnote-ref-3)
4. This is most obvious with respect to non-mentalist versions of SE, which hold that meanings are partly determined by altogether non-mental factors (cf. Speaks 2017). See footnote 21 for a brief comment about mentalist versions of SE such as Burge’s (1979) social externalism. Thanks to an anonymous referee for pushing me towards this clarification. [↑](#footnote-ref-4)
5. But see Horvath (2016) for an argument to the contrary. [↑](#footnote-ref-5)
6. See Burge (1986b, 1990) for similar proposals; see Lewis (1999), Reimer and Michaelson (2017) and Speaks (2017) for criticism. [↑](#footnote-ref-6)
7. I say ‘summarize’ because Cappelen does not really argue for these views. Instead, they are presupposed so as to build a theory of CE around them (Cappelen 2018, 63). [↑](#footnote-ref-7)
8. When untangling the elements of **Inscrutable – Lack of Control – Will Keep Trying**, Cappelen’s characterization of Will Keep Trying does not involve ‘should’ anymore. He merely says that we *will* and, perhaps, *must* keep trying to engage in CE. However, the original formulation of his **Inscrutable – Lack of Control – Will Keep Trying** thesis above as well as many other statements in the book make clear that he also believes that we *should* keep trying. See fn 12 for references. [↑](#footnote-ref-8)
9. Compare this line of thought to what Mark Richard (ms) writes about the ameliorative project: “It will be successfully carried off only if a large number of those who think about the world using C, expressing those thoughts with word W, come to do something we might call ‘changing their concept C’ in ways that reflect the revisionary analysis while continuing to use W to express their (revised) concept C.” See also Mikkola (2013). [↑](#footnote-ref-9)
10. Cappelen explicitly stresses the parallels between ideal theory and CE (83). [↑](#footnote-ref-10)
11. Ironically, in the eyes of many contemporary political philosophers, even Rawls’ use of ideal theory appears to be too detached from our actual non-ideal circumstances (Mills 2005, Sen 2006; see Valentini 2012 for a useful overview). It is therefore unlikely that Cappelen’s claim can be justified by an analogy to contemporary proposals in political philosophy. [↑](#footnote-ref-11)
12. Cappelen expresses this when he says that “purely descriptive philosophy must be abandoned” (Cappelen 2018, 47), or that the “[a]ssessment and improvement of concepts is at the core of philosophical practice, no matter what the topic” (Cappelen 2018, 47-8). It is really hard to see how to square these passages with the claim that CE is inscrutable and out of control. [↑](#footnote-ref-12)
13. How high the required degree of likelihood has to be in order to count as ‘sufficient’ is a question that I will not pursue in this paper. I suspect that this will partly depend on the context. For my purposes, all that matters is that a probability below 1 can be sufficient for control. [↑](#footnote-ref-13)
14. These definitions are slightly altered from Alston’s original proposal. Note also that both immediate and long-range control come in degrees, depending on the likelihood that, given the action(s) performed by *S*, *c* will come about. Note further that I take the notion of ‘being able to take a certain action/set of actions’ to come with an epistemic constraint: *S* must have some grounds for deciding which (course of) action is the one that will bring her the desired result. This epistemic constraint can be fleshed out in numerous ways, e.g. in terms of belief, justified belief or knowledge. For the purposes at hand, it won’t be necessary to give it a precise rendering. [↑](#footnote-ref-14)
15. Reference change in connection with the causal theory of reference has received some discussion in the literature as well. Gareth Evans uses the examples of ‘Madagascar’ as well as a fictional case about two switched babies to show that proper names can switch reference (cf. Evans 1973, 196). Michael Devitt (1981, 140 ff.) makes essentially the same point. LaPorte (2004) argues that causal theories are compatible with reference change for natural kind terms and offers extensive discussions of real life cases like ‘species’ or ‘jade’. In his view, many of the historical cases which externalists typically describe by saying that scientists have discovered the essence of a kind and thereby corrected the past usage of the term referring to the kind, really do involve refinements, and thus changes, of meaning. [↑](#footnote-ref-15)
16. LaPorte’s (2004, 151-154) discussion of the change from Euclidian to elliptic geometry is structurally analogous: “Surely the meaning of ‘straight’ changes from Euclid’s use to that of elliptic geometry…elliptic geometry changes the *paradigms* used to “*ground*” straightness. Extension is determined in part by paradigms: “*This* is straight, *that* is not” (pointing). And speakers counted circles around spheres “*non*straight”, at least in the contexts of interest here. Only with the advent of non-Euclidian geometry did the great circles on a sphere come to serve as examples of “straight lines”. [↑](#footnote-ref-16)
17. I have stated this explanation in Evans’ terms, but it can easily be restated in Devitt’s terms as well. [↑](#footnote-ref-17)
18. Mikkola (2009) suggests that there are some differences in extension nevertheless – the queen, for example, seems to be a female human being who is not oppressed in virtue of her observed female bodily features (565). [↑](#footnote-ref-18)
19. A similar suggestion is made by Thomasson (2007), chapter 2.3. See also LaPorte (2004), especially p. 6-7. [↑](#footnote-ref-19)
20. Although I have been focusing on causal theories of reference throughout this paper, it is likely that similar points can be made also about other versions of SE such as reference magnetism (e.g. Lewis 1999) or social externalism (e.g. Burge 1979). According to what Dorr and Hawthorne (2014) call ‘Lewis’ toy theory’, the best overall interpretation of a piece of language, and thus the meaning assigned to it, is the one that strikes the best overall balance between interpreting people as speaking truly and maximizing the naturalness of the assigned referents. Meaning change can thus be affected by working on the use part. Social externalism holds that meanings are partly determined by deference to experts. The idea that the use of experts plays a special role in determining reference does not change the fact that reference change supervenes on changes in the way people use language. [↑](#footnote-ref-20)