Conceptual Engineering: The Master Argument[1]

Herman Cappelen

Forthcoming in Burgess, Cappelen and Plunkett *Conceptual Ethics and Conceptual Engineering* (OUP 2018)

I call the activity of assessing and improving our representational devices 'conceptual engineering'. The aim of this paper is to present an argument for why conceptual engineering is important for all parts of philosophy (and, more generally, all inquiry). Part I of the paper provides some background and defines key terms. Part II presents the argument. Part III responds to seven objections. The replies also serve to develop the argument and clarify what conceptual engineering is.

[1] Much of the discussion in this paper draws on Cappelen (2018). This can be seen as an elaboration of an argument in Part II of that book, but this is the more refined version of the argument (better than the one I present in the book). I presented this paper at the Philosophy Mountain Workshop in Telluride and at the Arché Conceptual Engineering in St Andrews. I'm grateful to all the participants for helpful discussions and in particular to Alexi Burgess, David Plunkett, Kevin Scharp, and Seth Yalcin for detailed comments.

Part I: Background and Explanation of Central Terms

If we use ‘conceptual engineering’ as I suggested above, i.e. to mean *the project of assessing and improving our representational devices*, then if you think ‘concepts’ are the core representational devices, conceptual engineering amounts to the following: It is the project of assessing and then ameliorating our concepts.[2] For example:

- An epistemological conceptual engineer will assess epistemic concepts with the aim of improving them.
- A conceptual engineer in moral philosophy will aim to assess and improve our moral concepts.
- A metaphysical ameliorator will try to improve our core metaphysical concepts.
- A semantic ameliorator will try to improve our concepts for thinking about meaning and communication.

This normative project contrasts with a descriptive one. The descriptivist aims to describe the concepts we have—to describe our epistemic, moral, metaphysical, semantic, and so on, concepts. One important strand in the history of philosophy is a battle (or tension or at least division[3]) between Descriptivists and Revisionists. The distinction will not be simple or clear-cut and the battle lines have been drawn in different ways in different time periods. But in each time period and in all parts of philosophy, we find these two fundamentally conflicting attitudes or goals. For some, success is measured by a true description of, for example, what knowledge, belief, morality, representation, justice, or beauty is. For others, the aim is figuring out how we improve on what we have: how can we improve on (our concepts of) knowledge, justice, belief, beauty, etc.? Those with the former aim tend to find the latter unintelligible (or

[2] In Cappelen (2018), I end up not describing conceptual engineering in this way, but the argument in this paper can be presented independently of those reservations.

[3] How best to describe it is discussed further in connection with Objection (7) below.

naive) and those with the latter aim tend to find the former complacent, uninspired and lazy.

Nietzsche is perhaps the paradigm of a philosopher advocating what he calls ‘absolute skepticism’ towards our inherited concepts. In *The Will to Power*, he writes:

> Philosophers ... have trusted in concepts as completely as they have mistrusted the senses: they have not stopped to consider that concepts and words are our inheritance from ages in which thinking was very modest and unclear....What dawns on philosophers last of all: they must no longer accept concepts as a gift, nor merely purify and polish them, but first make and create them, present them and make them convincing. Hitherto one has generally trusted one's concepts as if they were a wonderful dowry from some sort of wonderland: but they are, after all, the inheritance from our most remote, most foolish as well as most intelligent ancestors. ...What is needed above all is an absolute skepticism toward all inherited concepts (Nietzsche 1901/68, section 409)

Strawson thought about the history of philosophy in part as a battle between revisionists, like Nietzsche, and what he called descriptivists. At the beginning of *Individuals* he distinguishes between *descriptive* and *revisionary* metaphysics:

> Descriptive metaphysics is content to describe the actual structure of our thought about the world, revisionary metaphysics is concerned to produce a better structure. (1959: 9)

One of his characterisations of the revisionists’ objection to descriptivists is acute. He imagines the revisionist insisting that metaphysics is

> … essentially an instrument of conceptual change, a means of furthering or registering new directions or styles of thought. (1959: 10)

In that brief introduction Strawson also gestures at a way of writing a history of philosophy based on the distinction between descriptivists and revisionists. He says

> [p]erhaps no actual metaphysician has ever been, both in intention and effect, wholly the one thing or the other. But we can distinguish broadly: Descartes, Leibniz, Berkeley are revisionary, Aristotle and Kant descriptive. Hume, the ironist of philosophy, is more difficult to place. He appears now under one aspect, now under another. (1959: 9)

This contrast is salient in many philosophical domains today. There are lively revisionist projects in moral philosophy, theories of truth and logic, feminist and race theory, and so on.[4] At the same time, there are descriptive projects dominating large swaths of philosophy. The paradigm might be epistemologists' endless efforts to correctly describe what the English word 'knows' means.[5] The same kind of descriptivist aim is found throughout philosophy: just think of the obsession within the philosophy of language with the minutiae of natural language and its semantics (for example the now-over-a-century-long debate about how the word 'the' works in English, or again the flurry of work on the English word 'if'), or efforts to describe our concepts of 'freedom', or 'self' or 'object'. These are more often than not pursued as purely descriptive enterprises. Success is measured by descriptive adequacy—not by answering the question *what* should *those words mean?* and *what* should *the relevant concepts be?*[6]

Part II: The Master Argument

The argument I'll be defending in what follows—'The Master Argument'—is in no way original to me. It is a line of thought that can be seen as motivating many revisionist

[4] The literature is large here. A representative selection is: Railton (1989, 1993) for moral philosophy, Scharp (2013) and Eklund (2014) for truth and logic, and Haslanger (2012) and Appiah (1996) for the philosophies of gender and race. See *Fixing Language* chapter 2 section 1 for a more complete list.

[5] What Robert Pasnau laments, in his 2013, as the degeneration of epistemology into lexicography.

[6] For a brief history of the tension between revisionist and descriptivists in 20th Century philosophy, see Cappelen 2018 chapter 2.

projects.[7] Nonetheless, the argument in full generality is often left implicit. The aim in what follows is to articulate the most *general* (not domain specific) version of the argument and then consider a range of objections to it.

The Master Argument

1. If W is a word[8] that has a meaning M, then there are many similar meanings, M_1,M_2,...,M_n, W could have.
2. We have no good reason to think that the meaning that W ended up with is the best meaning W could have: there will typically be indefinitely many alternative meanings that would be better meanings for W.
3. When we speak, think, and theorize it's important to make sure our words have as good meanings as possible.[9]
4. As a corollary: when doing philosophy, we should try to find good meanings for core philosophical terms and they will typically not be the meanings those words as a matter of fact have.
5. So no matter what topic a philosopher is concerned with, she should assess and ameliorate the meanings of central terms[10].

This is a controversial argument that builds in many assumptions. I take the project of outlining a general theory of conceptual engineering to be, in large part, the project of defending this kind of argument. In contrast, most work done on conceptual engineering is domain-specific and done by those working in those domains. For example, you might be interested in questions such as:

[7] See for example Eklund 2014, and Chalmers 2011, and the work in general of David Plunkett and Timothy Sundell (2013), and of Haslanger (many of the papers collected in her 2012).

[8] In this paper I don't individuate words semantically. If you like to use the word 'word' so that it denotes a lexical item that has its meaning essentially, then you can translate from your way of speaking to mine by substituting 'lexical item' for 'word' in this paper. The metaphysics of words is difficult and relevant, but not addressed in this paper (see Kaplan 1990 and Cappelen 1990 for some discussion).

[9] Or, as David Plunkett suggested to me, maybe 'good enough meanings' is good enough. Maybe aiming for 'best possible' is too ambitious. I'm open-minded on this issue. This is one of several places where one can articulate related versions of the Master Argument.

[10] As will become clear below, amelioration sometimes involves improving the meaning while keeping the lexical item fixed, and sometimes it involves the introduction of a new lexical item with an improved meaning.

- What should our moral concepts be like?
- What should our gender and race concepts be like?
- What should our concepts of, say, *civilian* or *person* or *family* be?

Those specific topics derive their significance from the importance of the specific subject matters addressed. The Master Argument, on the other hand, isn't domain-specific. This, I take it, shows that there is a *general* research project here, that is of interest independently of the case studies.

Replies to Seven Objections

In what follows I address Seven related objections to the Master Argument:

- Objection (1): Why think that if a word, W, has a meaning M, then there are many similar meanings W could have?
- Objection (2): In what sense can one meaning be better than another?
- Objection (3): Why not think the meanings words have are the best they can be (or at least very, very good)?
- Objection (4): If we change the meaning of an expression, won't that result in massive verbal disputes and changes of topic?
- Objection (5): Aren't meaning assignments normatively neutral, as long as each thing worth meaning is meant by some word or other? Some things are worth meaning, but why does it matter whether a given word means one of those things or something else?
- Objection (6): Why think the importance of the revisionist project undermines the importance of the descriptive project? Why think there's a tension between the two approaches? Aren't they complimentary?

- Objection (7): If we are to engage in conceptual engineering, don't we have to assume that meaning assignments are within our control? If they are out of our control, how can we meaningfully engage in conceptual engineering?

In answering these objections, I hope to a), clarify the nature of conceptual engineering, and b), outline the central challenges for conceptual engineering as a field.

Objection (1): Why think that if a word, W, has a meaning M, then there are many similar meanings W could have?

I respond to this objection in two steps: First I motivate the idea of similar meanings, and then the idea that W could have one of these similar meanings.

> *On the idea of similar meanings:* There's little agreement on what meanings are and the Master Argument is neutral on that foundational issue. However, it's hard to think of any account of what meanings are that's incompatible with this claim. A way to illustrate this is to think, in common with many philosophers, of meanings as at least intensions, i.e. functions from points of evaluation to extensions.[11] So 'freedom', 'knows', 'justice', 'belief' and all other expressions are associated with an intension and this intension is a function that gives a value (an extension) for each point of evaluation. Now, just change the function a little bit and you have a similar but different a meaning (if meanings either are or determine intensions). Suppose, for example, 'knows' picks out a relation between an agent and a content. The literature on the definition of 'knows' has given us literally hundreds of proposals for intensions of 'knows' that are very similar — they differ a tiny bit on how to deal with Gettier cases or lottery cases or some other weird scenarios, but are otherwise very similar. 'Knows'

[11] For the purposes of this paper, I will be neutral on what points of evaluation are—they might be worlds, world/time pairs or something more complicated. I don't need to take a stand on those issues in this paper (though I do have views, for which see Cappelen and Hawthorne 2009).

presumably picks out one of these functions, but there are very many similar ones.

On the idea that words can change from one meaning to another one: So far so good, but why think that once a meaning has been fixed, it is changeable? Why not think that meaning assignments get fixed and then are stuck eternally? First-pass answer: a word, say ‘marriage’ *could* mean anything. We could, right now, use it to mean what ‘camel’ or ‘soup’ does. There’s nothing in that sign that makes those meaning assignments impossible. If it’s possible, now, for ‘marriage’ to mean *camel*, then it could also mean one of the meanings that are similar to its current meaning.

In reply to this one might think that conceptual engineering is a matter of implementing gradual changes, not the kind of change exemplified by the possibility of assigning *camel* as the meaning of ‘marriage’. However, we know that there can be such gradual changes and that they happen constantly. Historical or diachronic linguistics is a field devoted entirely to the study of various forms of syntactic and semantic changes over time. These fields are in part the study of how meanings of words evolve gradually (from one meaning to a similar one) over time. The claim that gradual semantic change is impossible is refuted by the findings of these well-established research fields.

Two issues raised by this reply will be addressed later:

a) I just argued that meaning adjustment is possible. I didn’t say that it was easy or within our control. I return to the question of how we can (or cannot) be in control of meaning change in reply to objection 7 below.
b) I’ve focused on cases where we want to improve the meaning of a particular lexical item. In some cases, we don't care about the lexical item in question, but rather want to introduce a *new* lexical item for the improved or alternative meaning. There’s a range of options here and they’re discussed in reply to objection 5 below.

Objection (2): In what sense can one meaning be better than another?

There's an extensive philosophical tradition for thinking that the concepts or meanings our words express can be defective and can be improved along various dimensions, and so for thinking that some meanings can be better than others. This view can be found throughout the history of philosophy, but in what follows I'll focus on some versions found in the 20th and 21st centuries. Carnap's idea that intellectual work typically involves explication is a paradigm. The core thought was that the meanings we assigned to our words could be defective and for Carnap the central defects were indeterminacy and vagueness. Explication was a process of improvement. For Carnap, improvements were measured relativized to purposes. Here is how Anil Gupta explains the difference between an absolute and purpose-relative improvement:

> An explication aims to respect some central uses of a term but is stipulative on others. The explication may be offered as an absolute improvement of an existing, imperfect concept. Or, it may be offered as a 'good thing to mean' by the term in a specific context for a particular purpose. (Gupta 2015 §1.5)

I'd like to focus on the idea that an explication is a 'good thing to mean' by the term in a specific context for a particular purpose. Here is an illustration from Gupta:

> The truth-functional conditional provides another illustration of explication. This conditional differs from the ordinary conditional in some essential respects. Nevertheless, the truth-functional conditional can be put forward as an explication of the ordinary conditional *for certain purposes in certain contexts*. Whether the proposal is adequate depends crucially on the purposes and contexts in question. That the two conditionals differ in important, even essential, respects does not automatically disqualify the proposal. (Gupta 2015 §1.5)

Much the same idea of explication can be found in Quine's *Word and Object* (Quine 1960), where Quine writes about explication:

> We do not claim synonymy. We do not claim to make clear and explicit what the users of the unclear expression had unconsciously in mind all along. We do not expose hidden meanings, as the words ‘analysis’ and ‘explication’ would suggest; *we supply lacks*. We fix on the particular functions of the unclear expression that make it worth troubling about, and then devise a substitute, clear and couched in terms to our liking, that fills those functions. Beyond those conditions of partial agreement, dictated by our interests and purposes, any traits of the explicans come under the head of "don't-cares" (§38). Under this head we are free to allow the explicans. (Quine 1960: 258-9)

On this more general understanding of explication, there is no unique correct explication of any term, the improvement is relative to contextually specific purposes. With that in mind, there is no reason why there should be a fixed set of theoretical virtues that are used to measure improvement. In certain contexts, non-theoretical virtues/advantages could make a big difference.

So understood, much work in social and political philosophy can be seen as a continuation of Carnap’s proposal. Take, for example, Sally Haslanger’s work on gender and race concepts. An important element of that work is a proposal for how gender and race terms can be ameliorated, i.e. can be given better meanings. The dimensions of assessment that she has in mind are different from Carnap’s, but they are part of the same general project: of improving our concept along *many and diverse dimensions.*

A great deal of philosophy engages in this form of amelioration, often without making a big deal out of it. Consider for example Clark and Chalmers’ paper ‘The Extended Mind’ (1998). Clark and Chalmers propose, among other things, that ‘A believes that p’ be used in a way that makes it true even when p is a proposition that A has access to only with the assistance of various ‘external’ devices (Clark and Chalmers 1998). In connection with that proposal, they consider various objections of the form: *Well, that’s just not how we use ‘belief’ in English*. In response they say:

> We do not intend to debate what is standard usage; our broader point is that the notion of belief *ought* to be used so that Otto qualifies as having the belief in question. In all *important* respects, Otto's case is similar to a standard case of (non-occurrent) belief [...] By using the 'belief' notion in a wider way, it picks out something more akin to a natural kind. The notion becomes deeper and more unified, and is more useful in explanation. (Clark and Chalmers 1998: 14)

Clark and Chalmers' goal is not primarily to describe our current concept of belief—they want to *revise* our concept. Note two particularly important features of the proposal. (i) Their revision changes both the extension and the intension of the concept of 'belief'. (ii) They briefly provide a justification for the revision: it is, they say, more useful in explanations. Their new revised concept is also 'deeper' and 'more unified'. So our current notion is defective, as it is not sufficiently unified, not sufficiently deep, and not sufficiently useful in explanations. These kind of ameliorative moves are made in all parts of philosophy. For some salient recent examples think of Haslanger's work on race and gender concepts, Eklund and Scharp on truth, and Railton's proposed revision of core moral terms (see footnote 4 for references). For a wide range of illustrations, see Cappelen (2018 chapter 2), Plunkett and Sundell (ms), and Ludlow (2014). Beyond philosophy it's easy to find lively debates about (what at least looks like) meaning assignments. Consider discussions about what should be in the extension of 'torture', or 'person', or 'marriage', or 'rape'. These are not plausibly construed simply as efforts to find an descriptively adequate account of what these words actually mean. Imagine if a semantically omniscient god told us that 'torture' has a semantic value that doesn't include waterboarding. That's extremely unlikely to stop the debate over whether waterboarding is torture. One way to explain that is to construe it as a debate over whether 'torture' *should* have a meaning that makes it include waterboarding in its extension.

So far, I've simply pointed out that there are people who think meanings can be improved. It goes beyond the scope of this paper to assess each of these proposals. Suffice it to say that if you find one or more of those views plausible, you should be on board with a version of premise (2) in the Argument.

It is, however, worth considering the denial of the claim that meaning assignments can be assessed. This is the view that there's no normative dimension along which one meaning for W can be better than an alternative meaning. More generally:

***Complete Neutrality**:* Meaning assignments are always normatively neutral.

To assess Complete Neutrality, consider the fact that meaning assignments have a huge influence on *what* we can think about and *how* we can think about those features of the world that we can think about. Both of these have all kinds of effects on humans, both inter- and intra- personally. To hold that those effects cannot be assessed seems implausible for a range of simple reasons: If, for example, Fs are important to a group of agents, then it's good for them to be able to think about Fs and not good if they can't think about Fs. If, however, thinking, theorizing, or discussing Fs is unfortunate for those agents, then nevertheless having an expression that denotes Fs is worse than not having one.

Objection (3): Why not think the meanings words have are the best they can be (or need to be)?

I hope objection (3) sounds silly: It's implausible that a cultural artifact that's generated in a messy, largely incomprehensible way that's outside our control[12] should end up producing something we can't improve on. Everything we humans produce can be improved—why think meaning assignments are ideal (or good enough) straight off? Even if there are degrees of appropriateness, it would be amazing if we got exactly the right degree right off the bat.

A natural thought in this vicinity is nicely captured by Austin. In 'A Plea for Excuses' Austin says that "ordinary language . . . embodies . . . the inherited experience and acumen of many generations of men. . . . If a distinction works well for practical

[12] Or at least so I argue in reply to Objection (8) below.

purposes in ordinary life (no mean feat, for even ordinary life is full of hard cases), then there is sure to be something in it, it will not mark nothing” (Austin 1956: 11). Austin might be right: the carvings up that have survived over many generations are likely to ‘mark something’. But note that even if this is true, we’re not even close to the claim that there’s no room for improvement. The Austinian thought gives us reason to think we’re not totally wasting our time thinking with, say, the predicates we have, but in no way moves us towards the claim that they can’t or shouldn’t be improved. This is of course recognized by Austin who goes on to say that ‘...ordinary language is not the last word: in principle it can everywhere be supplemented and improved upon and superseded.’ (Austin 1956: 11). The challenge here is to recognize when ordinary language is good enough and when it can be improved upon. This is a deeply normative project, not primarily a descriptive one, and it is continuous with the kind of engineering projects described above. In the passage from *The Will to Power*, quoted above, Nietzsche points out that our concepts are ‘the inheritance from our most remote, most foolish as well as most intelligent ancestors.’ Given the impact of the most foolish, it would be naive in the extreme to trust our conceptual dowry in any domain.

Austin connects the survival of what he calls ‘a distinction’ to the promotion of certain purposes. Again, as Nietzsche points out, these purposes are typically remote and often the purposes of fools. As purposes change (maybe because some of us become less foolish) we need to adjust the way we carve things up. Sometimes this will be the result of moral and political evolutions. In other cases, the changes are driven by needs and purposes that result from technological changes. In yet other cases, the changes are the result of theoretical developments. As purposes change, meanings need to change as well. This is yet another reason to endorse an attitude closer to Nietzsche than to Austin: continuous and radical skepticism towards inherited concepts and distinctions.

Objection (4): If we change the meaning of an expression, won’t that result in massive verbal disputes and a change of topic?

Many philosophers have had concerns roughly of the following form. Suppose ‘F’ means M. We then ameliorate and in so doing revise the meaning of ‘F’. The following will be the result: Pre-ameliorators using ‘F’ will be talking about something other than those using ‘F’ post-amelioration. The result of the amelioration will be a change of topic. That, again, can lead to verbal disputes: The pre-ameliorator asserted ‘Fs are G’ and the post-ameliorators say ‘Fs are not G’. It looks like a disagreement, but if there’s been a change in meaning of ‘F’, then there’s no disagreement. Moreover, if people pre-amelioration had put massive effort into trying to answer the question: ‘Are Fs G?’, it now looks like we’ve lost track of that question. The question asked by post-ameliorators when they utter ‘Are Fs G?’ is a different question.

Worries of this sort constantly come up in discussions of revisionary traditions in philosophy. My favorite illustration is Strawson’s objection to Carnap’s account of explication (in his 1963) but we find the same kinds of concerns in the work of, for example, Haslanger and Railton.

This is a concern I take seriously and much of my *Fixing Language* is an effort to respond to Objection (4). Here is a summary of the reply:

> We know independently of considerations having to do with conceptual engineering that speakers who use the same sentence, S, with different but relevantly similar semantic contents, can use S to *say the same thing*. So suppose A and B both utter S, but the semantic value of S differs a bit in their two utterances (say, because they occupy slightly different contexts.) We can still, in many contexts, report them by saying that they have both said that S. That’s to say, we can use S to say what they have both said. So same-saying can be preserved across differences in semantic content. I suggest we use that as a model for what happens when meanings are ameliorated as described above. The result is a change in, at least, extensions and intensions, but that’s consistent with a preservation of same-saying. If it can preserve same-saying, there’s a sense in which they both talk about the same subject matter. I call this a

preservation of *topics*. So even though the meaning of, say, ‘family’ can change over time, we can still say that there’s continuity of topic among those who use ‘family’: they are talking about families. Evidence for this is that two speakers who utter ‘Families are G’, one pre-amelioration and one post-amelioration, can both be described as having said that families are G.

To endorse this you need to buy a collection of claims that that can be summarized as follows:

- First, you have to accept that same-saying can be preserved despite semantic differences. Same-saying data from speakers who use context sensitive terms is an important source of evidence here. An adjective like ‘smart’, for example, will fix a comparison class (or a cut off on a scale) in context. There can be two utterances of ‘Jill is smart’ where the comparison class or cut off differs a bit (i.e. they have different semantic contents), but where it is true to say that both A and B said that Jill is smart. So this is evidence that same-saying can be preserved across semantic difference. I think it’s fair to say that there’s a broad consensus about this and that it constitutes more or less common ground among many of those who think about meaning and communication. (For more on this, see Cappelen and Lepore 2005.)
- Second, you have to accept that this notion of same-saying across semantic difference can be used to establish a notion of topics and topic-preservation that covers what happens when we engage in conceptual engineering. Alternatively, you could take ‘sameness of topic’ as a primitive and use it to explain why we treat speakers as samesayers. Either way, I suggest we treat this cluster of concepts as basic and not aim for a reduction. The core question, for a theory of same-saying or topic-preservation, is what theoretical use they can be put to.

I explore these issues further in Part III of Fixing Language and my conjecture is that even those who oppose the Austerity Framework that I develop there can endorse this part of it.

Objection (5): Aren’t meaning assignments normatively neutral, as long as each thing worth meaning is meant by some word or other?[13]

Some things are worth meaning, but it doesn’t matter whether a given word means one of those things or something else. Suppose we’ve established the meaning of a word, W, is defective along some important dimension and we come up with an ameliorative strategy. A natural question is: when implementing this strategy, why keep using W? Since you’re introducing a new (allegedly improved) meaning, why not use a new word to mark the change? This is an important question in many cases and to help articulate it, I distinguish between *lexical expansion* and *lexical improvement:*

> *Lexical Expansion*: Where a new meaning is introduced as the meaning of a new expression.
> *Lexical Improvement*: where a new meaning replaces the meaning of an already ‘in use’ lexical item has.

The objection say: We have no reason to prefer Improvement over Expansion. The underlying thesis is Pro-Expansion:

> *Always-Expand:* New meanings should always be attached to new lexical items.

One way to motivate Always-Expand is through thinking about a strategy involved in what Chalmers (2011) calls ‘the subscript gambit’. According to Chalmers, most philosophical concepts are surrounded by similar concepts which constitute clusters. For example, there’s a bunch of concepts in the vicinity of the concept of freedom. The English word ‘free’ might pick out one of those, but we should not think that the concept our word happens to pick out is particularly interesting or useful. What we should do

[13] This useful articulation of the objection is due to Alexis Burgess. A discussion with Cian Dorr also helped crystallize this reply.

instead is explore the entire conceptual neighborhood, and then introduce a range of new expressions. Chalmers tends to describe these using subscripts, so we have ‘freedom$_1$’, freedom$_2$’,..., freedom$_n$’. We should keep each of these in our conceptual arsenal, so to speak. They can be useful for different purposes—and having all of them lexicalized enables us to express a wider range of truths.

Chalmers thinks philosophers have spent too much time (thinking they are) fighting over what freedom *really* is by asking the question “Which one of freedom$_1$, freedom$_2$,...,freedom$_n$ is *really* freedom?” (although they wouldn’t phrase it that way, of course). As he sees it, that’s a fight over what the semantic value of ‘freedom’ *simplicter* is, and that is an uninteresting question. Chalmers’s thought is similar to the one in the Master Argument: The English word ‘free’ has ended up with a particular semantic value, but it could easily have had any of a bunch of similar meanings—that it ended up with the particular meaning it has is in part the result of random and intellectually insignificant factors. Chalmers (2011) says that obsessing over what that semantic value is can only be motivated by a ‘fetishistic’ value system.

With that as a background, the motivation for Objection (5) and Always-Expand should be clear: just as it would be irrational to care about which freedom-concept is assigned to ‘freedom$_1$’ and which to ‘freedom$_2$’, it doesn't matter which freedom concept is assigned to ‘freedom’. What matters is that we articulate and lexicalize the range of interesting concepts in this vicinity.

Before turning to my reply, a couple of initial remarks:

- First notice that a proponent of Always-Expand is deeply involved in the core activities of conceptual engineering: Evaluating meanings and reflecting on ameliorative strategies. She is not opposed to conceptual engineering, but is making a particular proposal for how ameliorative strategies should be implemented.
- Second, I agree with the objection that the issues raised here are entirely non-trivial: terminology is important. Choice of lexical items is important. If Always-Expand is true, that’s a deep fact about the practice of conceptual engineering. In

other words, I don’t think this is just a trivial issue about words, because issues about words are hardly ever trivial.

That said, I think Always-Expand is false: Sometimes it is important to preserve the lexical item. Here are four considerations against Always-Expand:

1. Sometimes we care about lexical effects

The lexical item *itself* can have cognitive and non-cognitive effects on us that we want to preserve. In *Fixing Language* (chapter 11), I call these ‘lexical effects.’ Here are some illustrations:

- Brand names: There is a reason why companies spend an enormous amount of resources protecting their brand names. 80-90% of the value of the Coca-Cola company lies in its ownership of the name ‘Coca-Cola’.[14] What does that mean? It is of course complicated, but one thing it means is that if Coca Cola had to change the name of its core product, then the value of the company as a whole would decline dramatically. What’s important for our purposes is this: in a scenario where the company was not allowed to put words ‘Coca-Cola’ on their product, people’s propensity for buying and consuming the product would decline. This proves that choice of lexical item matters. A proponent of Always-Expand wouldn’t be able to convince Coca-Cola executives that the name of their product is irrelevant.
- Names of children: That words’ lexical properties affect people in interesting ways is shown by a study on how political affiliation influences the way parents name their children. As surprising as it might (or, on further reflection, might not) seem, there is a correlation between what a child is named and the political affiliation of the parents who name it. In particular, a study suggested that parents in liberal neighbourhoods are more likely to opt for 'soft' letters in naming

[14] The economist Aswath Damodaran gives an analysis of the value of the Coca Cola brand name in a blog post at http://aswathdamodaran.blogspot.com/2013/10/the-brand-name-advantage-valuable.html.

their child, such as 'l's and 'm's, while conservative people are more likely to opt for harder sounds like 'k' or 't'. Thus baby Liam is more likely to be the product of liberals while Kurt might follow in his parents' footsteps and become conservative.[15]

The kind of signaling that is involved in name choices is, again, not about the meaning of the name (or the meaning of sentences containing the name). It is about triggering certain kinds of lexical effects and what the study shows is that parents choices are guided by lexical effects even when they are not aware of it.

- Finally, some of the debate over the term ‘marriage’ illustrates the point. The word ‘marriage’ has a certain effect on people and in the debate over same-sex marriage it was important for proponents that the lexical item ‘marriage’ was used about their relationship. To see that, note that the following wouldn’t suffice to meet the demands of all proponents of same-sex marriage: a proposal to introduce another term—say, ‘zwagglebuggle’—that denotes the same rights and obligations as ‘marriage’ and same-sex couples could say that they were ‘zwagglebuggled’, but weren’t entitled to use the term ‘marriage’ about their relationship. One reason why some proponents of same-sex marriage would reject this is that the lexical item ‘marriage’ has important cognitive and non-cognitive effects and those are important in the debate over ‘same-sex marriage’. The aim is, at least in part, to change of meaning of ‘marriage’.[16,17]

These are not isolated examples. As Chalmers point out:

> Ideal agents might be unaffected by which terms are used for which concepts, but for nonideal agents such as ourselves, the accepted meaning for a key term will make a difference to which concepts are highlighted, which questions can

[15] https://www.livescience.com/37196-politics-baby-names.html

[16] I’m here assuming, for illustrative purposes, that this involves a meaning change. It’s an open question whether it does.

[17] This is a point Chalmers agrees with: he is clear that in some—maybe many—cases, words matter (see the discussion of ‘torture’ in his 2011). Chalmers himself isn’t a proponent of Always-Expand (his view was introduced because parts of his view could be (ab)used by someone sympathetic to that thesis.)

> easily be raised, and which associations and inferences are naturally made.(Chalmers 2010: 542)

Many words have a massive effect in social, political, legal, medical, and inter-personal contexts. We have no reason to think theoretical contexts are immune to these kinds of effects. In cases where preservation of lexical effects are important, amelioration will involve changing the meaning of the current word, not introducing a new word. I think lexical effects are ubiquitous and of enormous importance to communication. They are poorly understood and under-investigated (chapter 11 of Fixing Language provides the beginning of a theory.)

2. The original lexical item as marker of topic continuity

In reply to Objection 4 I outlined an account of topic continuity across semantic changes. I'll use 'freedom' as an example. Suppose that at time t 'freedom' denotes a certain property, P. Then a semantic change happens, and as a result, at time t_1, 'freedom' denotes a different property, P*. This does not prevent there from being continuity of topic in uses of 'freedom' at t and t_1. Speakers who, at t, utter 'Freedom is G' and speakers who, at t_1, utter 'freedom is G' can *say the same thing.* They can all be talking about freedom and say about it that it is G. The topic—freedom—can be preserved through semantic changes. That is an important kind of continuity in discourse. This continuity is why we can say about speakers at t and t_1 that they are trying to answer the same questions and it's what underpins their agreement. Now suppose we left the word 'freedom' behind and instead introduced indefinitely many new words 'barakuns','hostomas', 'notacabil', etc., for each new meaning. In that case we would lose an important marker of discourse continuity. The connection to previous discourse would disappear. Continuity of lexical item is an important marker of topic continuity.

In response to what I said above, a proponent of Always-Expand could say:

> You say continuity of lexical item is important, but is it essential? It might be no more than a contingent heuristic—useful for restricted agents like humans, but in principle dispensable. Ideally we should just name all the different properties in the freedom neighborhood, and then some of these (or many of them) would constitute topic continuity. An ideal agent wouldn't need lexical continuity to recognize that continuity.

In reply, I would point out that the communicative features that are important for non-ideal agents matter to us. Humans are very much non-ideal and lexical continuity is what we—with our limited minds and fragile access to semantic and communicative content—often need to track topic continuity. Maybe it's not needed by gods or by massively improved humans but that does not make it less important for us (who are not gods or cognitively enhanced). Here is one way to see why we need it: We care about inter-contextual and inter-conversational continuity. We can individuate conversations so that they last over long time, take place in different places, and involve broad range of people who will never know one another. People who will never meet each other can participate in a conversation about, say, democracy and terrorism. Technology has changed the paradigm of a conversation: it's not longer a group of people standing within hearing distance of each other, it is, rather, people transmitting data across the internet to people they might never encounter. Since words don't come with little definitions attached or lines that mark topic continuity, we often have to use lexical continuity as markers of topic continuity. What ties conversations about democracy or terrorism together is often 'democracy' and 'terrorism'. If someone started using 'swugleding' for an ameliorated meaning of 'terrorism', that would most likely fail to connect to the continuous conversations about freedom. There are millions of existing tokens of 'terrorism' and the best way for you to connect to those is to use 'terrorism'. 'Swugleding' likely won't do the work, no matter how carefully introduced. Even 'terrorism*' will most likely fail because hardly anyone will have access to the meaning of the '*'.[18]

[18] Note: I am not saying that it is always important to preserve lexical item. The claim is that lexical preservation often matters.

3. The anchoring role of the original lexical item

To see what I have in mind, consider again the subscript strategy. It start with freedom and then finds properties in the vicinity (or neighborhood) of it. The subscript strategy presupposes that there is a freedom *cluster*. It presupposes that some properties are within the freedom cluster and other properties are not. For example, the property of being one of my eyes isn't in the freedom cluster (speaking in the subscript lingo: it's not a candidate for being one of 'freedom$_1$', 'freedom$_2$', etc.). That's because it's not one of the properties in the neighborhood of freedom. The metaphors of 'neighborhood' and 'clusters' play important roles in the description just given and we need an account of what demarcates the clusters or neighborhoods. I tacitly introduced such a criterion (in connection with the example of freedom): it has to be appropriately/relevantly related to freedom (without a subscript). We have no other place to start. When theorizing we start with freedom (or, more generally, one of the non-subscripted lexical elements) and then we find properties that are related to it. If that's the general strategy, we need 'freedom' simpliciter as the anchor point for amelioration. It is what topic continuity (being about freedom) is measured relative to. If something like that is right, then an appeal to freedom simpliciter might be theoretically *indispensable* if you want to preserve topic continuity.[19,20]

4. The Role of Lexical Items in Social Ontology

According to many views of social ontology, language play an important role in the creation and preservations of social facts.[21] There isn't much agreement on just what that role is, but there is fairly broad agreement that they play a role. Much here is going

[19] None of this is to deny that you could just ignore topic continuity and describe properties independently of what 'clusters' or 'neighborhoods' they belong to. Nor is it to deny that having created the cluster, we couldn't then shift emphasis to one of the others (i.e., the anchoring doesn't have to be continuous).
[20] See Chapter 17 of Fixing Language for more on this line of thought.
[21] For a paradigm of the kind of view I have in mind, see Searle 1995, chapter 3. See also section 4.6 of Brian Epstein's Stanford Encyclopedia entry on Social Ontology (Epstein 2018)) and the references in that entry.

to depend on the exact role language plays and which part of language is most important and settling these issues goes beyond the scope of this paper. With all those reservations in place, there is an important conditional claim:

> **Possible Connection Between Lexical Items and Social Facts:** If lexical items play some role in the creation and preservation of social facts, then changing the meaning of a lexical item might contribute to a change in social reality.

Illustration: Suppose a term like 'family' plays a role in creating and sustaining the social category of families. If so, then improving the meaning of 'family' can be contribute to a change and maybe also an improvement in that part of social reality - it can change/ameliorate the social category of families. Suppose, instead, we introduce a new term, 'scramies' with the improved meaning. That might not have the same effect.

This point is made with a lot of 'might's and 'maybe's. We don't yet know enough about the role of language (and expressions in particular) to make confident claims here. It's an important issue that conceptual engineers should explore further.

Objection (6): Why think the importance of the revisionist project undermines the importance of the descriptive project? Why think there's a tension between the two approaches? Aren't they complimentary?

The Objection: In Part I of this paper I described a *tension* between a descriptive project and a revisionist project. But why aren't they complimentary projects? You can do some describing and some amelioration, and these go hand in hand. Moreover, I have described conceptual engineering as a two-step process: first describe deficiencies and then develop ameliorative strategies. That first step is at least in part descriptive,[22] so the revisionist project presupposes the descriptive project. Setting it up as a conflict is misleading.

[22] That it's a defect is of course a normative judgement, but presumably that judgment relies on a description of what it is like (or what properties it has).

Before replying to this objection, I should note that it contains an important element of truth. There’s no inconsistency between the two projects; they could complement each other. Despite this, I think it’s accurate to talk in somewhat loose terms about a ‘tension’ between the two projects. Here is why:

Reply 1: The goals and purposes you have when doing research guides much of what you do. For example, consider someone interested in how cells grow and divide as a purely intrinsically interesting topic. Their research will be guided in very different directions compared to someone who is interested in understanding cancer and is doing the research in order to develop a cure for it. There’s obviously no incompatibility between the two projects, but the focus will be immensely influenced by the goal. For a closer-to-home example, consider someone interested simply in the syntax and semantics of ‘true’ compared to someone interested in ‘true’ because of the semantic paradoxes. This will give rise to very different research projects and there’s a tension in this sense: a lot of what the one is doing will be irrelevant to the other because their research direction and priorities will differ significantly. Or compare the following two: on the one hand, someone interested in the semantics and syntax of terms like ‘woman’, and on the other, someone like Sally Haslanger who is interested in how such terms ought to be used to classify group for political purposes. Again, there’s no incompatibility between the two projects, but the differences in goals will lead to difference in priorities and this again will shape radically different research projects. However, none of this is to say that these differences in goals, priorities and direction imply that the ameliorator shouldn’t in part be guided and restrained by descriptive insights.

Reply 2: What is impossible—or at least incompatible with the Master Argument—is to be a ‘pure’ descriptivist. A descriptivist who claims to have no interest in or need for conceptual engineering shows a lack of understanding of the Master Argument. A corollary of that argument is that all the concepts involved in describing conceptual engineering should themselves be subject to critical assessment. So should the concepts used to describe and execute the descriptive project. We have to assess and improve on the following concepts: ‘concept’, ‘conceptual defect’, ‘descriptive work’, and so on. In other words, the very terminology with which you engage in the descriptive project is itself subject to assessment. So a ‘pure’ descriptivist would show a lack of understanding for the need to assess and improve on the concepts used to engage in the descriptive project (and even the concepts involved in describing the contrast between the descriptive and the amelorative project).

Objection (7): If we are to engage in conceptual engineering, don't we have to assume that meaning assignments are within our control? If they are out of our control, how can we meaningfully engage in conceptual engineering?

The objection: Setting up the objection will be slightly more work than in the previous cases. Here is the basic thought: Conceptual engineering, I claim, involves identifying representational defects and then finding ameliorative strategies. This seems to imply (or presuppose) that successful conceptual engineering requires that meaning assignments are in large part within our control. Why construct ameliorative strategies if we can’t implement them?

One reason I take this objection very seriously is that I think meaning assignments are in large part incomprehensible and outside human control. In what follows I first explain why, and then reply to the objection.

Background for the objection: The metasemantic facts are out of control and inscrutable. I focus on what two broad kinds of metasemantic theories tell us about meaning determination: externalist theories and internalist theories. On the standard externalist story, content determining factors include:

- Introductory events that typically happened long time ago and were performed by people we don’t know anything about. The introductory events will be of two broad kinds: demonstrative (“let ‘F’ denote those kinds of things”) or descriptive ("let 'F' denote the things that are G"). Note that if this is right, then in many cases the facts surrounding these introductory events will be unknowable to us. We have no way to access what happened: there will be no written record and, absent a time machine, we can't know what happened. We don't know what was pointed to and we don't know the details of the descriptions used. Moreover, we don't know the motivations for these introductions.

- Externalist theories also appeal to chains of reference transition. In such communicative chains, expressions are 'passed along' from one speaker to another and there is some kind of reference-preserving element in the communicative chain. Sometimes those chains are not reference preserving; in such cases, we get reference shifts.
- Other externalists, such as Timothy Williamson, talk more generally of meaning as supervening on use patterns over time where this connection between use and meaning (or reference) is chaotic in the sense that there's no algorithm that takes us from use patterns to meanings. The connection is too complex—indeed it might be in principle too complex for humans to grasp. We can't know about all the particular uses, and even if we did know about them, the way in which meaning is generated by such use is too complex for the human mind to grasp.

An important feature of these kinds of views is that we are in large part *not* in control of the reference determining process. We're not, for example, in control of what happened far into the past, we're not in control of what happened in the transition periods. In general, no one is in control of the total pattern of use. Nor are we in control of the supervenience relation that takes us from complex use-patterns to meanings.

It's tempting to think that if your metasemantics is more internalistic, meanings would be more within our control and so meaning change would be easier. Burgess and Plunkett, for example, say:

> The textbook externalist thinks that our social and natural environments serve as heavy anchors, so to speak, for the interpretation of our individual thought and talk. The internalist, by contrast, grants us a greater degree of conceptual

> autonomy. One salient upshot of this disagreement is that 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. (Burgess and Plunkett 2013: 1096)

I agree with one part of this: from the point of view of an internalist theory, meaning change and revision is possible. That's all that's needed to support the second premise in the Master Argument. However, and this will be relevant later, I disagree with the claim that internalism, as such, puts us more in control of these changes. Internalism is a supervenience claim: the extensions and intensions of expressions supervene on individuals (and then lots of bells and whistles to elaborate on this in various ways, but the bells and whistles don't matter right now). Suppose the meaning of my words supervene in that way on *me*. Note first that this is compatible with the meanings and extensions supervening on features of me that I have no control over. It's also compatible with it being unsettled and unstable what combination of internal features ground reference. So there's just no step from Internalism to control. Moreover, even if meaning supervenes on something internal that I have control over, control doesn't follow: it could supervene on something we could control, but the determination relation from the supervenience base to meanings/extension could still be out of our control. For example, even if there's supervenience on what we want or intend or decide, the supervenience relation doesn't have to make it the case that semantic values are what we intend for them to be, what we want them to be or what we agree on them to be (for all we know, it could be a total mess or get us to the opposite of what we want, intend or decide).

In sum: both externalist and internalist theories makes meaning change possible, though both of those theories make it hard to see how this is a process we can be in control of. This is why objection (7) is pressing.

Reply: In summary form, my reply is that conceptual engineering shares this feature with most normative theorizing. On the view I defend, the tools we think with are often defective, but there’s very little we can do about it. We can talk and think about it, but doing so has hardly any effect. Compare that to me talking to you about crime in Baltimore, poverty in Bangladesh, or the Trump presidency. Such talk is unlikely to have any effect on what happens in Baltimore, Bangladesh, or Trump. The ineffectiveness of talking is an almost universal aspect of large-scale normative reflections. Anyone who spends time thinking and talking about large-scale normative matters should do so without holding out too much hope that their talking and thinking will have significant or predictable effects on the relevant aspect of the world. If you think your views and theories about crime in Baltimore, poverty in Bangladesh, or the Trump presidency will have a significant or predictable effect on either, you’re extremely likely to be disappointed (and to end up feeling you’ve wasted the part of your life that has been devoted to these issues). There are of course small-scale local issues where normative reflections will have a direct effect. If I think my daughter shouldn’t have an ice cream, then, at least in a few cases, the result will be that she eats no ice cream. Moving to slightly larger-scale issues—say speed bumps in the street where I live—my opinions, views, and pleadings will have tiny effects, but already these effects will be fairly marginal, unsystematic, and unpredictable (as I’ve discovered). On the view proposed in Cappelen (2018), changes in extensions and intensions of words are far over on the large-scale and unpredictable side. Much closer to crime in Baltimore than to speed bumps in Sofies Gate.[23] So, in sum, the worry that I’ve painted too bleak a picture of the prospects of conceptual engineering simply fails to take into account the relevant comparison class. What I say about conceptual engineering shouldn’t be surprising and doesn’t make the activity of trying to engineer concepts much different from a wide range of other human efforts to think about how things should be.

Conclusion

The Master Argument provides a general argument for the importance of conceptual engineering. It has nothing to say about *particular* deficiencies or ameliorative

[23] And as I just said, even the speed bumps turned out to be more or less completely out of my control and I ended up concluding that this particular instance of ‘local’ activism was a waste of time. The conclusion is not that we shouldn’t do it, but rather that if we do it, we should do so without illusions.

strategies. As a heuristic it’s useful to think of conceptual engineering as having two parts: the general theory and the specific applications. We should expect a two-way interaction: the general theory will inform the specific cases and the specific cases will inform the general theory. It should also be clear from the discussion above that there can be many frameworks for thinking about conceptual engineering. What one takes conceptual engineering to be (both when thinking about the general theory and about specific cases) will be shaped in large part by what one takes meanings and concepts to be, what one assumes about metasemantics, and what one takes to be conceptual defects and virtues. One advantage of the Master Argument is that it is neutral on those questions and so can provide a kind of common ground for all those who see conceptual engineering as central to philosophy.

Bibliography

Appiah, Kwame Anthony (1992). *In My Father’s House: Africa in the Philosophy of Culture*. Oxford University Press.

Austin, John (1956). A plea for excuses. *Proceedings of the Aristotelian Society* 57: 1–30.

Cappelen, Herman (1999). Intentions in words. *Noûs* 33: 92–102.

Cappelen, Herman. (2018). *Fixing Language*. OUP.

Cappelen, Herman, and Lepore, Ernest (2005). *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Wiley-Blackwell.

Cappelen, Herman & Hawthorne, John (2009). *Relativism and Monadic Truth*. Oxford University Press UK.

Chalmers, David J. (2011). Verbal disputes. *Philosophical Review* 120 (4): 515–66.

Clark, Andy, and Chalmers, David J. (1998). The extended mind. *Analysis* 58 (1): 7–19.

Dorr, Cian & Hawthorne, John (2014). Semantic Plasticity and Speech Reports. *Philosophical Review* 123 (3):281-338.

Eklund, Matti (2014). Replacing truth? In Alexis Burgess and Brett Sherman (eds.), *Metasemantics:New Essays on the Foundations of Meaning*. Oxford University Press, 293–310.

Epstein, Brian (2018). Social Ontology. *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/social-ontology/>.

Gupta, Anil (2015). Definitions. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2015 edition), <http://plato.stanford.edu/archives/sum2015/ entries/definitions/>.

Haslanger, Sally (2012). *Resisting Reality: Social Construction and Social Critique*. Oxford University Press.

Kaplan, David (1990). Words. *Proceedings of the Aristotelian Society*, Supplementary Volumes 64: 93–119.

Ludlow, Peter (2014). *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford University Press.

Nietzsche, Friedrich. (1901/68). *The Will to Power*. Translated by W. Kaufmann. New York City: Random House.

Pasnau, R. (2013). Epistemology idealized. *Mind* 122 (488): 987–1021.

Plunkett, David, and Sundell, Timothy (2013). Disagreement and the semantics of normative and evaluative terms. *Philosophers' Imprint* 13 (23): 1–37.

Plunkett, David, and Sundell, Timothy (ms). Work on conceptual engineering, title to be determined.

Quine, W. V. (1960). *Word and Object*. MIT Press.

Railton, Peter (1989). Naturalism and prescriptivity. *Social Philosophy and Policy* 7 (1): 151.

Railton, Peter (1993). Noncognitivism about rationality: benefits, costs, and an alternative. *Philosophical Issues* 4: 36–51.

Scharp, Kevin (2013a). *Replacing Truth*. Oxford University Press

Searle, John (1995). *The Construction of Social Reality*. Free Press.

Strawson, P. F. (1959). *Individuals*. Routledge.

Strawson, Peter F. (1963). Carnap's views on conceptual systems versus natural languages in analytic philosophy. In Paul Arthur Schilpp (ed.), *The Philosophy of Rudolf Carnap*. Open Court, 503–18.