Amelioration v. Perversion

Teresa Marques

Abstract

Words change meaning, usually in unpredictable ways. But some words' meanings are revised intentionally. Revisionary projects are normally put forward in the service of some purpose – some serve specific goals of inquiry, and others serve ethical, political or social aims. Revisionist projects can ameliorate meanings, but they can also pervert. In this paper, I want to draw attention to the dangers of meaning perversions, and argue that the self-declared goodness of a revisionist project doesn't suffice to avoid meaning perversions. The road to Hell, or to horrors on Earth, is paved with good intentions. Finally and more importantly, I want to demarcate what meaning perversions are. This, I hope, can help us assess the moral and political legitimacy of revisionary projects.

1 Introduction

Word meanings change. In a paper in this volume, Glasgow (2019) nicely illustrates a variety of changes in word meaning and reference:

> 'Sick' used to just mean ill; now it—like 'bad'—also means good. It seems we can now use 'literally' to mean figuratively. And then there are terms whose old meaning or referent is simply lost and taken over by a new one—outright replacement. Famously, 'Madagascar had been the name for part of the Somali peninsula, not that big island off of Africa's eastern shore. But Marco Polo got the location wrong. When map-makers subsequently doubled down on his mistake, the referent of that word underwent its change, and we now know Madagascar to be the island, not the peninsular region (Burgess 2014). This instance of referential replacement is hardly unique. 'Awful' once meant inspiring awe. It now instead means really bad. Even our hallowed example of 'bachelor' meant young knight before it referred to a marital status, pausing along the way to acquire rank of university achievement of a certain kind. 'Fantastic' (from imaginary to wonderful), 'silly' (from worthy to foolish), 'senile' (from the senescent generally to those with dementia specifically), 'tool,' and on and on: language is littered with conceptual change

These examples illustrate the malleability of words' meanings. Theories about the meaning of words for social roles and categories raise interconnected questions. One is metaphysical:

What is the nature of the social things we talk about? Another is conceptual: How do we think about the social world, and how does our thinking relate to the nature of social reality? And how should it relate? The latter question concerns so-called ameliorative projects. Such theories aim to answer other additional questions, for instance, what is the point of having the concept in question? What concept (if any) would better serve our social or political goals? And who are *we*?[1]

In this paper, I argue that we should add one more question to this list, namely: *How can we assess the legitimacy of ameliorative projects?* I will try to answer this question by drawing from deeply problematic historical cases. Based on these cases, I'll introduce two notions of *meaning perversion*. This, I believe, will put us in a better position to understand the limits of the moral and political legitimacy of ameliorative projects. Permissible meaning revisions are those that *are not* meaning perversions. This still says nothing about which revisions *we should* engage in, i.e., those that are required. But it gives us a way of circumscribing the meaning revisions we shouldn't pursue.

Suppose we are interested in understanding crucial concepts of the language of justice and politics, e.g., in understanding the concepts expressed by words like 'democracy', 'freedom', 'fair elections', 'citizen', 'the people', etc. Sally Haslanger has persuasively made the case that we can mean different things by "conceptual analysis", and that our reflection on the role of the language of politics can take different approaches. In a 2006 article, Haslanger distinguishes between *manifest*, *operative*, and *target* concepts. A manifest concept is the concept one thinks guides one's categorizing, whereas the operative concept is the concept that corresponds to actual categorization patterns. The *target* concept is the concept that, *all things considered*, we should employ *given our interests, the facts*, etc (Haslanger, 2006, 99). It is the concept that, in the end of the *ameliorative* project, we should be using.

The question I address in this paper – *how do we assess the moral and political legitimacy of revisionary projects?* – is thus essential to know if our projects are indeed *ameliorative*. Haslanger considers the question of how to know if an analysis is ameliorative, and says "whether or not an analysis is an improvement on existing meanings depends on the purposes of the inquiry." (Haslanger, 2012, fn 1, p. 367) But I take it that it would be too naïve to say that an analysis is an amelioration just in case the "engineers" of the revision have positive social or political purposes. Some features of what Maynard and Benesch (2016) call Dangerous Speech, which I'll introduce below, show that people can be convinced of the

[1]These questions are illustrated in debates about race, for instance in Appiah (2006), Glasgow (2003), Andreasen (2000, 2005), Kitcher (2007), Haslanger (2003, 2006), Machery et al. (2009), Diaz-Leon (2015).

goodness of a purpose while bringing out very bad outcomes. And hence, whether a revision is ultimately an amelioration cannot depend exclusively on how convinced theorists, or activists, are of the presumed goodness of their ends.

I take it, also, that there is some tension between *the concept we should be using, all things considered* and *the concept we should be using, given our aims, interests, facts*. The intersection of what we should do, all things considered, and what we should do, given our aims, etc., leaves open the possibility that the answer to the legitimacy question may be epistemically unavailable, or that we can remain wilfully ignorant of such an answer. It would be desirable to have some constraints or guidelines on how to assess a revisionary projects' legitimacy.

Now, when theorists consider *conceptual engineering* or *conceptual ethics* projects, they often are concerned with the concepts our words should express, given the specific aims of a theory. Theoretical concerns are normally removed from the political domain. For instance, Tarski (1943) considered that 'truth' as used in natural languages is an incoherent notion, and that it gives rise to paradox. He argued that 'truth' should be defined in a way that would allow it to play its foundational role in a semantic theory, while insulating the theory from the contradictions and paradoxes that the natural language use gives rise to. We could say that, for the purposes of the theory, a revision of the meaning of 'true' is viable when the new meaning fulfills the aims set by the theory.

We can think that similar theoretical goals can be pursued in social and political domains. Yet, there are differences between theoretical domains that do not touch on socially relevant notions, and theoretical domains that promote meaning revisions which will transpire into language use 'in the wild'. Whereas the notion of truth in Tarski's theory is not meant to replace the use of 'truth' in informal natural language use, meaning revisions about social kinds or social relations often have both theoretic and social aims. Redefining what 'marriage' means in actual legislations has changed the extension of 'marriage' in the real world.[2]

This paper differs from other recent articles that try to pin down normative constraints on conceptual engineering. For instance, Simion (2017) is optimistic about the prospects of conceptual ameliorations, but introduces only epistemic constraints. As she argues, a concept should be ameliorated only insofar as this does not translate into epistemic loss. More recently, Podosky (2018) argues *contra* Simion that epistemic loss is not an adequate criterion, but rather that conceptual revision should be allowed whenever it has the "capacity to causally

[2]I'm grateful to Pablo Rychter and Esa Díaz León for discussion about this point.

influence the world". Moreover, he suggests, normative questions should address *feasibility* constraints. He says that the claim that a concept should be engineered is right if

> ... its desired causal influence on social reality is feasible; and the claim that a concept should be engineered is wrong if its desired causal influence on social reality is infeasible. Importantly, the notion of feasibility mentioned here is epistemic: It is about the feasibility of representational accuracy.(Podosky, 2018, 13)

In other words, the engineering of a concept is feasible only if it is possible for that concept to come to accurately represent (social) reality. This would answer the normative question *when should a concept be ameliorated*. For instance, the revision of 'marriage' and the progressive change of legislation in liberal democracies around the world is feasible in Podosky's sense, and this answers the question of whether 'marriage' should be ameliorated.

This is *not* what I have in mind when I raise the question of the legitimacy of conceptual revision.[3] What I have in mind is the assessment of the moral and political legitimacy of revisionist projects. A project *could* well be feasible in Podosky's sense, and still be illegitimate in the senses that I articulate in this paper. That is not the case of the amelioration of 'marriage' – its revision is not only feasible, it also has largely improved the lives of same-sex couples by extending to them rights that other couples already enjoyed (and it removed no one else's rights). But there may be problematic revisions. Another way of characterizing what I have in mind is to ask when are revisionist projects not ameliorative but perversive.

I will make the case that there is no obvious answer as to what 'the value of a meaning revision' can mean beyond serving specific theoretical purposes. I also argue that it is hard to balance the value of pursuing theoretical aims against the possible harms caused by using the revised words 'in the wild', so to speak. I'll argue that the conviction that an analysis is pursued with good intentions does not guarantee the goodness of the outcome. And even when revisionary projects achieve those aims, the aims themselves may not be desirable. On the contrary, we as theorists should be cautious of deep feelings about the desirability of our purposes. There may be also cases where there is no answer to the question of what we should do. Given this, it would be desirable if we could have some guidelines on how to assess the moral legitimacy of a revisionist project. It would be useful to know, for instance, which are the projects that *we should not pursue*. To this end, I'll consider two alternative ways of

[3]Lawford-Smith (2013) offers a more engaging discussion of what political feasibility may mean. She says that feasibility is often used for ruling out political theories that can't be implemented, but that it should rather be used as a tool for ranking alternative theories "along one of the dimensions relevant to making decisions about what to actually do" (Lawford-Smith, 2013, 245). She also says that feasibility is independent of both desirability and risk. Hence, although understanding feasibility as a way to understand conditional probability of different political theories to meet certain ends is an essential step in assessing the political projects we can engage in, it does not assess the desirability of the ends or the risks themselves.

understanding what meaning perversions are.

In section 2, I start by summarizing some lessons about meaning perversions from history, and then I point to some cognitive and affective biases that are conducive to condoning harmful practices and behavior, while protecting moral self-righteous convictions. In section 3, I extract consequences for ameliorative projects. In section 4, I try to characterize the conditions for legitimate ameliorative projects, and to define meaning perversions. Ultimately, meaning (or conceptual) perversions are projects that are politically or morally illegitimate. I will offer two ways of characterizing meaning perversions: those that can be characterized in terms of their *harmful effects*, and those that can be characterized in *constitutive terms*. Being alert to the possibility of meaning perversions (in either sense) should be a constraint on any project of conceptual or meaning revision.

2 Meaning revisions in the wild

2.1 Lessons from the Past and the Present

In LTI, *The Language of the Third Reich*, Victor Klemperer offers a chilling description of the corrupting power of language. Klemperer witnessed how, under the Third Reich, "huge number of concepts and feelings" were corrupted and perverted. This was the case of words like 'heroic', 'heroism', or 'fanatic'. In his diaries, he registered how the use of these words shifted in a way that indicated that, for instance, one could not be a hero unless one were a fanatic. At the same time, fanaticism was no longer regarded as a negative trait. Klemperer reported that even after WW2 ended,

> young people in all their innocence, and despite a sincere effort to eliminate the errors in their neglected education, cling to Nazi thought processes... as soon as this concept (*heroism*) was touched upon, everything became blurred in the fog of Nazism... and then replaced it with 'fanatical'. (Klemperer, 2000, 14)

He continues by drawing an illustrative analogy, saying that Nazi language "commandeers for the party that which was previously common property and steps words, groups of words, and sentence structures, in poison."

The journalist Masha Gessen has also written on the corrupting power of discourse. In her article from 2017, "The Autocrat's Language", she warns us of the damage Trump's public discourse does to social and political reality. In the article, Gessen draws a parallel between Trump's discourse, the discourse of autocrats in the former Soviet Union, and presently in

Russia. She says,

> A Russian poet named Sergei Gandlevsky once said that in the late Soviet period he became obsessed with hardware-store nomenclature. He loved the word secateurs, for example. Garden shears, that is. Secateurs is a great word. It has a shape. It has weight. It has a function. It is not ambiguous. It is also not a hammer, a rake, or a plow. It is not even scissors. In a world where words were constantly used to mean their opposite, being able to call secateurs "secateurs"—and nothing else—was freedom.
>
> "Freedom," on the other hand, was, as you know, slavery. That's Orwell's 1984. And it is also the USSR, a country that had "laws," a "constitution," and even "elections," also known as the "free expression of citizen will." The elections, which were mandatory, involved showing up at the so-called polling place, receiving a pre-filled ballot—each office had one name matched to it—and depositing it in the ballot box, out in the open. Again, this was called the "free expression of citizen will." There was nothing free about it, it did not constitute expression, it had no relationship to citizenship or will because it granted the subject no agency. Calling this ritual either an "election" or the "free expression of citizen will" had a dual effect: it eviscerated the words "election," "free," "expression," "citizen," and "will," and it also left the thing itself undescribed. When something cannot be described, it does not become a fact of shared reality. Hundreds of millions of Soviet citizens had an experience of the thing that could not be described, but I would argue that they did not share that experience, because they had no language for doing so. At the same time, an experience that could be accurately described as, say, an "election," or "free," had been preemptively discredited because those words had been used to denote something entirely different. Gessen (2017)

Klemperer talks of 'that which was previously common property', and Gessen of a lost 'shared reality'. This is, I think, one of the harmful effects of meaning perversions, but I will not elaborate further in this paper on what the loss of a shared reality or common property amounts to, and how it opens the way for autocracy. I call these uses of "freedom", "free expression of citizen will", "election", "heroic", "fanatical", etc. *meaning perversions*. I will consider two non-synonymous and (possibly) non-coextensional senses of *meaning perversion* in section 3.2: a causal consequentialist sense, and a constitutive sense. The next subsection illustrates the extent of the consequences of our cognitive failures and biases in some extreme cases.

2.2 The road to Hell is paved with good intentions

The end aim of meaning revisions can be self-interested. Some politicians' aim in using words like "law and order" or "justice" is not to promote the rule of law in the service of the interests of justice and fairness, but to use the justice system to protect themselves, or consolidate their power. Other politicians's aim in using words like "democracy" and "free elections" is not to allow for governments to be representative of their constituents, with the guaranteed

protections of citizen's rights and the rule of law, but to perpetuate their hold on power. Yet, meaning revisions often do have genuine social or political justice aims, or at least, aims that their proponents *believe* will bring about a better world. This section discusses some historical cases where extremely harmful consequences were seen as morally permissible, or even morally and politically required. People's inability to foresee those consequences are due to cognitive limitations that are common. Any of us can fall into the trap of motivated reasoning, for instance.

How does this relate to revisionist projects? We as revisionists may genuinely *believe* in the goodness of our purposes and methods. But there is a gap between what may be the best course of action to achieve a certain end, and our epistemic capacity in assessing those means to achieve that end. Moreover, the fact that we may desire certain ends does not make those ends desirable (all things considered), or desirable independently of the means that are set to achieve them.

Here is an extreme example of discourse that, from the perspective of the speaker and the target audience, is perceived as morally legitimate. In recent work, Maynard and Benesch (2016) discuss the conditions under which so-called *Dangerous Speech* can occur. They characterize Dangerous Speech as speech acts that are capable of encouraging approval of violence by an audience.

Maynard and Benesch (2016) argue that Dangerous Speech is a product both of the *context* and of the *content* of discourse, and that these feed into and overlap with each other. Inflammatory speech produced in a context where it cannot be disseminated (because the audience strongly disapproves of it or the speaker has no influence) does not amount to dangerous speech. I.e., in contexts where there is *no uptake* (and presumably, *no context update*), inflammatory speech does not amount to Dangerous Speech. The specific features of the *contexts* of Dangerous Speech are the *speaker*, the *audience* in its the socio-historical environment, and the availability of *means of dissemination* (Maynard and Benesch, 2016, 77).

If a community relies mainly on one source of information, the message spread by that source is more influential. But an influential or authoritative speaker addressing a volatile audience through mass means of information is not producing Dangerous Speech if the content is not inflammatory or hateful. I.e., unless the content can persuade the audience that violence is permissible or *morally justified by the circumstances*, speech is not dangerous. This does not require that the whole audience comes to engage in violent actions; just that the audience comes to condone acts of violence.

Dangerous Speech often occurs in social and historical contexts that increase the likelihood that the audience accepts that violence against certain people is morally permissible. It can be seen as rightful punishment for presumed past crimes, or as means to prevent presumed future (existential) threats. Socio-historical contexts of Dangerous Speech may include longstanding *grievances*, *resentment*, the *memory of historic injustices* (real or imagined), a weak or dysfunctional *justice system*, *competition for resources*, or *land disputes*. Influential speakers may manipulate and exacerbate the resentment against members of another group to capitalize on populist rhetoric and take advantage of those grievances for political gains.

The *correlation* between Dangerous Speech and mass violence has been established in various studies. In Rwanda, the station RTLM was the main source of the inflammatory messages. Yanagizawa-Drott (2014)'s statistical study of the effects of the virulent propaganda RTLM in Rwanda indicates that killings were 65-77% *higher* in Rwandan villages that received the RTLM signal, compared with those that did not (for reasons like topography) receive the signal. In a recent study, Müller and Schwarz (2018) discovered a correlation between Facebook use and an increase in anti-refugee attacks in Germany. In particular, they show that right-wing anti-refugee sentiment on Facebook predicts violent crimes against refugees in German municipalities with higher social media usage.

Maynard and Benesch (2016) offer a characterization of six possible features of the *content* of Dangerous Speech. These features need not all be present, and need not be present in the same way in all cases. The first is *dehumanization*, a notion that Tirrell (2012), Jeshion (2016), and Snyder (2017b) also rely on. These are forms of discourse that can do direct harm by the offense, denigration, or derogation of members of a target group as undeserving of the duties owed to them qua persons. The second feature of the content of dangerous speech is *guilt attribution*: members of a group are said to be guilty (as members of the group) of past crimes, e.g., rape or murder, stealing, responsible for current difficulties, destruction of the economy, occupation, or oppression, etc., and their guilt is offered as the moral justification for feelings of resentment and retributive action.

The third possible feature of the content is *threat construction*: an in-group accuses an out-group of being a (existential) threat, which again contributes to morally justify acting in self-defense. The fourth feature is the *destruction of alternatives* and the representation of a course of action as a historical necessity, or of alternative courses of action as impractical or inefficacious. For instance, Figes (2002) describes how citizens of the Soviet Union thought that the violence of the Stalinist era was the only possible and necessary path to Communism,

reporting someone that said: "I had my doubts about the Five Year Plan... but I justified it by the conviction that we were building something great... a new society that could not have been built by voluntary means". (Figes, 2002, 111)

The fifth feature is what Maynard and Benesch (2016) call *VirtueTalk*, through which the audience is motivated by "deep and unreflected feelings that something feels 'good' or 'bad', in particular feelings that induce positive moral self-appraisal", a "satisfactory mental image of themselves... often shaped by notions of ideal group-identities, that produces considerable self-esteem" (Maynard and Benesch, 2016, 84). It is a feature that, again, contributes to the moral justification of actions against an out-group.

The final feature of the content of dangerous speech is *Future-bias*, i.e., the promise of future goods. Future-bias is presumed to outweigh the short term difficulties the audience may have to endure, or the moral costs of the violence against others. Maynard and Benesch (2016) illustrate this again with examples from the former Soviet Union:

> But the anticipated benefits can also be extravagant and utopian—promises that a positive transformation of society will be brought about through a temporary violent transition, or that national unity and prosperity for a long-mistreated people can be obtained. In light of the expectation that Soviet violence would protect the revolution and usher in Communist utopia, Lenin assured his followers that in the future "the cruelty of our lives, imposed by circumstance, will be understood and pardoned. Everything will be understood, everything." Soviet ideology and the justification of massive violence and cruelty in the name of a promised future society of abundance convinced millions in the Stalinist era. The novelist Boris Pasternak wrote in a letter in 1935: "The fact is, the longer I live the more firmly I believe in what is being done, despite everything. Much of it strikes one as being savage [yet] the people have never before looked so far ahead, and with such a sense of self-esteem, and with such fine motives, and for such vital and clear-headed reasons."(Maynard and Benesch, 2016, 85-86)

Dangerous Speech arises when a perfect storm of cognitive biases and contextual conditions come together. But what is it that distinguishes forms of speech that are *dangerous* from those that are really *morally justified*? Suppose that it is true that a certain target group is guilty of past offenses, or presents a serious threat, and that one is persuaded that there are no viable alternatives to violent action. If this were true, one could be convinced that there are good reasons to act violently in self-defense, or retributively. Indeed, the great motivational strength of Dangerous Speech – its nearly irresistible pull – comes from its reliance on moral reasoning and motivation. But we should be cautious here. Resorting to violence in self-defense may be justified in specific cases. However, even if self-defense is justified, retributive violence may not be, as in fact I think it is not. A violent response to past offenses does not undue the offense or the suffering it caused, it can do little to prevent future harms, and is arguably

harmful in itself. To quote the Argentinian cartoonist Quino in one of the strips of his comic *Mafalda*, if all the bad people were killed, only the murderers would remain. And in fact, generally dangerous speech is not based on such solid reasons such as legitimate self-defense. Rather, it is motivated by attitudes that can easily go wrong.

The motivational force of dangerous speech depends on the reactive attitudes involved. The notion of reactive attitudes was introduced by P. F. Strawson (2008), and it includes attitudes like gratitude, resentment, contempt. These are attitudes that we have in reaction to another person's action towards us. They are essential both in interpersonal relations and in our moral lives.[4] Reactive attitudes like resentment or contempt guide our behaviour towards other people; they are "morally laden and broadly juridical or legalistic reactions to other presumptively reasonable people" (Manne, 2017, 197). Negative reactive attitudes are ways of putting down, punishing, diminishing another person, presenting her as deserving of the treatment (Jeshion, 2016, 91-4). Yet, our cognitive failures and biases – in ascribing guilt to others, in ruling out alternative courses of action, in seeing ourselves as virtuous (as *just knowing* that we're right), or in biases towards some future ideal, show that reactive attitudes and the actions they motivate can be misguided and morally unjustified.

Meaning perversions are not a minor issue, I believe. Through meaning perversions, as Victor Klemperer put it, "language does not simply write and think for me, it also increasingly dictates my feelings and governs my entire spiritual being the more unquestioningly and unconsciously I abandon myself to it". Meaning perversions played a part in making it possible for Soviet citizens to tolerate and condone the actions that brought about the hunger of millions of people, for instance the Holomodor in Ukraine (a fact many still deny or minimize).[5] What are the consequences of the cognitive failures of meaning perversions for revisionist projects? And how should we characterize meaning perversions? The next section tries to answer these questions.

3 The Limits of Ameliorative Projects

In this section, I will suggest that we, as theorists, should be defensive pessimists[6] when seeking to answer the question about the legitimacy of meaning revisions, i.e., about how to assess

[4]The notion has been deployed in recent moral and political philosophy, for instance by Manne (2017), Björnsson and Hess (2017), Goldman (2014), or Couto (2019).

[5]See Snyder (2017a).

[6]The notion of defensive pessimism was introduced in Norem (2001). It is a cognitive strategy that helps people to take proactive behavior to counteract possible negative outcomes.

the value and desirability of a meaning revision to a certain end. Indeed, we should not only be aware that any possible meaning revision may, and likely will, have harmful unintended consequences (as most of our actions do). We should also be aware that our cognitive failures often prevent us from seeing the harm our plans and actions bring about. Insofar as we can, we should minimize such harm.

We should approach revisionary projects with defensive pessimism, but also with honest humility. One way of deploying defensive pessimism, i.e., of taking proactive behavior to counteract possible negative outcomes, is to adopt "reflective equilibrium" as a method for finding coherence among our diverse considered judgments, bringing them into relations of mutual support and explanation.[7] The process should be understood as allowing for the revision of our convictions, given the facts and the foreseeable outcomes of our actions, a revision that honest humility can facilitate.

In a recent lecture, Haslanger (2018) argued against *ideal theory* and in favor of critical non-ideal theory, to decide how to answer questions in social ontology – what possible and actual scenarios are relevant to consider. This decision, she argued, depends on what we want to know and for what purposes. Haslanger claimed that to address concrete justice issues, for instance "How should we revise the educational system in Boston to be more fair?", it is not helpful to start from idealized examples and ideal theory:

> We can learn a lot about justice from considering concrete social circumstances. I would argue that such "bottom up" (v. top-down) theorizing is the best way to learn about justice, for our background presuppositions are tested against the messy reality we are trying to address.(Haslanger, 2018, 2)

I don't have a definite view here, but I am skeptical of the possibility of doing away entirely with ideal theory, and relying only on "bottom up" theorizing. After all, without gaining distance from immediate practical concerns (interests and purposes) people may never move away from the deep feelings that something feels right or wrong, and great harm can be caused by acting only on such feelings, as the previous section tried to show.[8] In the case of the justice of rightness of a revision of meaning, relying on bottom up theorizing, and on the feeling that something feels rights, can't suffice to guarantee the moral or political legitimacy of a revisionary project.

Perhaps 'the value of a meaning revision' can refer to its positive overall consequences, all things considered. The distinction between an amelioration and a perversion would then

[7] After Rawls (1971).

[8] In any case, following a reflective equilibrium approach can be pursued in a nonideal way. Ssee for instance Stemplowska and Swift (2012).

depend on the balance of good *versus* harmful consequences of a revisionist project. But how can a theorist setting out to advance a new categorization know that good consequences will outweigh the harmful ones?

In the opening section, I distinguished between meaning revisions that are pursued to meet specific theoretical aims, and those that are intended to have real social consequences. This is a distinction that has been made in discussions about the usefulness of the concept of race, for instance. In this context, Kitcher (2007) has defended a pragmatic biological view about race. He argued that there are nondenumerably many ways to sort people into biological categories. These possible divisions, he claimed, depend on our cognitive capacities and our purposes. Other authors like Andreasen (2004) have argued for a cladistic conception of human races (see also Andreasen (1998, 2000, 2005) and this volume). One could perhaps ameliorate the concept of race to meet the purposes of our theories about human biology, and define races as "ancestor-descendant sequences of breeding populations that share a common origin."(Andreasen, 2004, 425)

Yet, eliminativists like Appiah (for instance in his contribution in Wilkins et al. (1998)) argued that there are no human races because ordinary racial categories do not track any actual natural differences between people. Moreover, the harms caused by the false belief in a unique biological base for racial categories is an additional reason to abandon all talk of human races. In European Portuguese, for instance, the term 'raça' is conatively loaded. Most people on the left avoid using the word. In fact, the end of the fascist dictatorship in Portugal in the 1970s led also to the independence of former Portuguese colonies in Africa (the revolution to end the dictatorship was carried out by the military who wanted to end the war in the former colonies, which had been going on for approximately 13 years). In the new legislation drafted by a mostly socialist and social-democratic parliament, all uses of 'raça' were removed. A similar process led to the removal of mentions of 'Rasse' from German legislation:

> Recent debates about *Weisssein* ("whiteness") in Germany provide further evidence that recognition of social realities of white supremacy does not presuppose an account of race in terms of these realities. For example, Amjahid's (2017) book *Unter Weissen* ("Among Whites") dissects German practices of white supremacy but explicitly avoids realist appeals to *Rasse* by pointing out the "historical burden" (2017, 49) of the concept. In this sense, the German concept of race may indeed be more adequately understood in analogy to other failed concepts such as *witch*. While alleged witches were forced in very real social positions, claims about the reality of races in Germany seem just as misleading as claims about the reality of witches.(Ludwig, 2018, 8)

In reply to concerns like these, Kitcher says that the usefulness of racial categories will

depend on the theoretical purposes that are best served by having those categories. But, as he puts it,

> [E]ven if the concept of race plays a role in some lines of biological inquiry, the values of those lines of inquiry, and of pursuing them through retention of the concept of human race, would have to be sufficiently great to outweigh the potential damage caused by deploying this concept in the other contexts in which it plays so prominent a role, namely in our social discussions." (Kitcher, 2007, 302)

Kitcher's suggestion – that we can ameliorate the concept of race when the value of pursuing biological lines of inquiry is sufficiently great to outweigh the potential damage caused by deploying the concept in social contexts – seems a bit cavalier. The theorist is assuming that the value of his inquiry justifies not only categorizing people into clades, i.e., ancestor-descendant sequences of breeding populations with a common ancestor, which of course is useful for the theory. The theorist is, on assumption, also using the word 'race' to refer to a clade knowing that there have been millions of people who have suffered because of their categorization under different "races", and that talking of biological races will continue to feed the social discrimination perpetrated on this basis. Redefining the meaning of race terms for the important goals of the theory does not do away, or minimize, the negative impact the use of 'race' has in society.[9]

What can we learn from this discussion? Could we say that the Germans and the Portuguese have ameliorated their concepts of race by revising the terms used as empty, and hence removing them from legislation? I don't think that would be correct. Learning that a term has no referent is not the same as revising its meaning. 'Unicorn' does not refer to anything because there are no unicorns, not because people ameliorated the meaning of the word to give it a null-extension. Pointing out that there are species of mammals that may have given

[9]I don't address social constructionist non-eliminativist view were, except briefly in a paragraph below. There are several reasons for this. First, I don't want to occupy too much space debating race, which is not the focus of this paper. I thought that it was enough to focus on what (I'd say) the majority of humanity who thinks there are human races believes they are – biological kinds. Another reason for not engaging with non-eliminativist social constructionism is that I myself am now not clear about what to think of it, in spite of having had sympathies for the view. I suspect that some of the motivation for the social constructionist view of race arises from a "North-American-centric perspective", given the centrality and importance of racial relations in structuring North-American society. But in the rest of the world, although *racism* exists, it's not clear to me that the most appropriate way of addressing social inequality is on the basis of a non-eliminativist notion of race. Races are not obviously social constructions, unlike property, presidents, judges, professors, stock exchanges, etc. To say that race is socially constructed is a theoretical perspective that social theorists can adopt to explain both the fact that it is not biologically real and the fact that there are social relations (in several countries) that seem to be based on races. But the point of treating race as a social construction is both theoretical (something a theorist is of course entitled to do, pretty much like Tarski is entitled to redefine a notion of truth in the theory), and practical if the theorist suggests this concept is to replace the real world concept. But this is what I'm again skeptical about. How do we know that deploying the theoretical notion of race in the real world would help, e.g., in giving people fairer more just lives? For all we know, it may be that reifying social identities can be an obstacle to create common ground. So, perhaps we don't really need to move the theoretical notion "into the wild" to advance social justice.

origin to the concept of unicorn – horses and narvals –, and that the study of these organisms is useful in biology, does not entail in any way that biology would benefit from using the term 'unicorn' to refer to either species.

The hypothesis that opened this subsection was that a meaning revision could be an amelioration only if the values served by the revision outweighed the potential damage caused by that revision. But this isn't sufficient to justify carrying out a revision that would give an empty term a new extension, given the probability of causing actual damage by saying that the term has *any* extension. Even if such a meaning revision could serve certain values, and that is calculated to outweigh the risks, we should be aware that the *risks* may include not only the higher or lower probability of harm. We should also be aware of the higher or lower intensity or impact of the damage caused. Low probability-high impact risks, i.e., damages that are *more serious* even if not highly probable, should guide strict constraints on revisionist projects. It would have to be the case that there are no alternative terms that can play the required role in a theory without the potential of damage.

In the case of 'race', science can continue to research human reproduction and migration without using race-talk, as it can without using 'unicorn'. Race-talk has had, and continues to have, hugely harmful consequences. In the German or Portuguese context, the usefulness of pursuing lines of inquiry about human clades would be better served by calling them by the technical term, 'clades', and not by the loaded terms 'raça' or 'Rasse'. Given the prevalence of racism elsewhere, I'd risk suggesting that removing talk of race from legislation would be an improvement. As Ludwig (2018) suggests, issues related to justice can be addressed instead by talking of racialized practices and conceptions. This is also, in fact, the strategy that Robin Andreasen argues for in this volume. She distinguishes between pluralists and eliminativists about 'race'. She further distinguishes between selective race eliminativists (who hold that 'race' should be used only in social theory domains but not all, particularly not in biology) and reconstructive race eliminitavists (who argue that in all contexts 'race' should be replaced with other terminology). She argues that ultimately we should endorse reconstructive race eliminativism, fundamentally for moral reasons and concerns about the possible harmful effects of its use.

Now, even if it were true that there is only one way of classifying humans into different clades, it would still be the case that those divisions should not warrant racially based discrimination. For any biologically real characteristic, as for instance biological sex is, we should refrain from discriminating anyone on its basis. I take discrimination to differ from mere dif-

ferential treatment.[10] This is a different issue than the race problem discussed above. In the race case, we have good reasons to say that there are no races because the underlying biological reality does not provide a coherent way of matching real biological categories to pre-theoretical race conceptions, and race words are normatively loaded in highly pernicious ways. But we do track other biological categories fairly well, as is the case with sex categories, or conditions like albinism. Female people, or albino people, have suffered discrimination and harm for being what they are. That harm that does not justify doing away with categorizing sexually male and sexually female organisms, any more than the harm that albino organisms suffer (they are prime targets for predators, for a start) does not justify doing away with the distinction between albinos and animals with (some) melatonin. Doing away with the categorization does not do away with the harm. When those organisms are human, recognizing people for what they are may indeed be the required differential treatment that is necessary to address the injuries they are more prone to suffer. For instance, recognizing the differences between female and male symptoms of cardiovascular disease is necessary for properly diagnosing and treating females under 40 who may die from heart attacks or strokes (Stamp (2018)).

In recent work, Haslanger (2019) offers a more detailed answer to the question of how to assess the adequacy of ameliorative projects (I regret that I will not be able to do justice to the complexity of Haslanger's discussion here). Haslanger draws from work by Knobe et al. (2013) and Knobe and Prasada (2011) on dual character concepts, concepts that have both a descriptive and a normative dimension, where dual character concepts

> . . . are represented by (a) a set of concrete features and (b) a set of abstract values that the concrete features are seen as realizing. These two representations are intrinsically related, but they are nonetheless distinct, and they can sometimes yield opposing verdicts about whether a particular object counts as a category member or not. (Knobe and Prasada, 2011, 2965)

The dual character content view offers a promising prospect for understanding the complexity of meaning change, and it is fundamental to understand meaning perversions, at least in the one of its senses, as I will argue in the next sections.

To summarize, it's still unclear what 'the value of a meaning revision' can mean beyond serving theoretical purposes, and how the value of pursuing theoretical aims can be balanced against the possible harm caused by using the word in the wild. Recall that, according to Haslanger (2006), an *ameliorative* approach to the language of politics itself will ask which concepts we *should* use, given our aims. But as we've seen, the conviction that an analysis

[10] See Lippert-Rasmussen (2013) for a discussion on ways to define discrimination.

is pursued with good intentions does not guarantee the goodness of the outcome. And even if a project achieves those aims, the aims themselves may not be desirable. On the contrary, theorists should be cautious of "deep feelings" about the desirability of their projects and aims. There may be also cases where there is no answer to the question of what we should do. Given this, it would be desirable to have some constraints or guidelines on how to assess the moral legitimacy of a revisionist project. This is what I try to do in the next section. I will take advantage of the difference between harmful perlocutionary effects and harmful illocutionary effects, but do that end I will introduce my preferred way of discriminating between the different elements that play a role in updates to a conversational common ground.

4 The Legitimacy of Ameliorative Projects

In this section, I introduce some theoretical resources that can help us understand how language can motivate people to act, and to have either illocutionary or perlocutionary effects that contribute to structure social relations. These resources are drawn not only from philosophy of language, but also from philosophy of action and of emotions. I will start by briefly describing how context can be updated with speech performed under different illocutionary modes. I will then make use of these resources to distinguish between two senses of meaning perversion. The notions I will introduce, in particular those of expressive presuppositions, will play a crucial role in the explanation of words that have expressive normative or evaluative presuppositions as part of their meaning. I will here assume that the notion that Haslanger makes use of, that of dual-character concepts, can be reinterpreted as multidimensional aspects of meaning, and that efforts to change one dimension of meaning may not be accompanied by changes in the other dimension.

4.1 Common ground and motivational set

In recent joint work, García-Carpintero and I propose an expressive presuppositional account of derogatory language. I think that our view can be extended beyond derogatory language to other kinds of discourse that encode normative or evaluative expressive presuppositions. On our view, expressive presuppositions are not just propositions to be added to a common ground as shared beliefs. Expressive language, e.g. that involving slurs, includes a cognitive (i.e., descriptive), and a conative dimension. Literal uses of pejoratives and slurs make requirements on a shared conative record, governed by *sui generis* norms specific to affective attitudes and

their public manifestations (Marques and García-Carpintero (2019)).[11]

We believe that rethinking the nature of the expressive conative dimension of language use requires conversational contexts to have illocutionary structure, including the different classes of contents to which speakers are committed in different modes: in the way we are committed to our *beliefs*, but also in the way we are committed to our *intentions*, to our *affective attitudes*, and to the *questions guiding our inquiries*. In felicitous contexts, these different commitments are mutually shared, and license presuppositions. As Stalnaker (1978)'s account of assertion emphasizes, an accepted assertion comes to be presupposed afterwards, allowing for the satisfaction of presuppositional requirements later on in the discourse. Similarly, accepted questions under discussion (QUDs), directives, and expressives can come to be taken for granted, constraining the legitimate moves that can later be made.

Our account fills in some of the details of the proposal made by Langton (2012):

> I want to propose, in an exploratory spirit, the idea that the phenomenon of accommodation might extend beyond belief—beyond conversational score, and common ground, as originally conceived—to include accommodation of other attitudes, including desire and hatred. My remarks here will inevitably be programmatic. But to convey the general idea: just as a hearer's belief can spring into being, after the speaker presupposes that belief, so too a hearer's desire can spring into being, after the speaker presupposes the hearer's desire; and so too a hearer's hatred can spring into being, after the speaker presupposes that hatred. Stalnaker's common ground can perhaps be extended to include not just common beliefs, and other belief-like attitudes, but common desires, and common feelings, as well. Speakers invite hearers not only to join in a shared belief world, but also a shared desire world, and a shared hate world.(Langton, 2012, 140)

The context update made with the literal use of a slur is an update not only on our shared cognitive common ground – the set of propositions that come to be commonly accepted under a belief mode–, but also on our shared "motivational set" – the set of intentions, evaluative dispositions, or desires that are shared under *intentional* and a *sentimental modes*.

The suggestion adapts Bernard Williams (1979)'s notion of a 'motivational set' – the set of dispositional attitudes, plans, intentions, emotions, etc., that identify the reasons agents have for acting, and are an integral part of their *practical reasoning*. A motivational set can be generalized to be part of *common* ground, which would include those motivational attitudes that are in fact common, and known to be such. And it can be generalized to be part of the 'motivational score', i.e. those attitudes that are part of the conversational score. This distinction may be useful. People often accept to follow norms that they don't agree with

[11] Cepollaro and Stojanovic (2016) also defend a hybrid evaluative account of pejoratives and thick evaluative terms, where the evaluative presupposition expresses that the referent of the term is good or bad in some way.

or that they disavow, and accept to act in accordance with widely shared and permissible evaluative attitudes that they do not actually have. But under certain conditions, social norms can be eroded or changed.[12] When there's conflict between different norms, plans, or evaluative dispositions, they can't be jointly implemented, followed, or satisfied. If sufficient people accept a norm, when another incompatible one is already in place, then this acceptance can *erode* the preexisting norm.

People may have doubts about the prospects of shared emotions under a *sentimental mode*, even if the prospect of shared plans does not raise the same doubts.[13] But authors like Salmela and Michiru (2016) have outlined an account of collective emotions that links the intentional structure of joint actions and underlying cognitive and affective mechanisms. Collective emotions can function as both motivating and justifying reasons for jointly intentional actions, in some cases even without prior joint intentions of the participants. And Salmela and von Scheve (2017) deploy research on collective emotions to explain right-wing radical populism, illustrating the usefulness of a notion of a common motivational set to explain the functioning of Dangerous Speech. They argue that there are two psychological mechanisms underlying the rise of right-wing populism: *ressentiment* and *emotional distancing*. Ressentiment explains how negative emotions like fear and insecurity transform shame into anger towards "our" perceived "enemies". Emotional distancing separates the social identities that inflict shame and other negative emotions from "us".

I suggest that our understanding of revisionary projects should not only incorporate the new developments on the dual character of concepts into our understanding of ameliorative projects, as Haslanger proposes to do. It should also incorporate work on the hybrid nature of evaluative and normative language, and on recent research on the role of emotions in social processes.[14] If words's meanings can serve to make not only assertive illocutionary acts. where we describe things as having specific features, but make also expressive speech acts, where the things (situations, events, or people) we describe are taken for granted as realizing certain norms of values through the emotional or affective attitudes we express.

[12]Bicchieri (2005, 2017) work models how shared social norms that are followed in a society rely on people's normative and empirical expectations with respect to what others will do and feel, and the conditions under which normative change occurs.

[13]On shared plans, see for instance Kutz (2000). Researchers like Charlow (2016), Portner (2016), and Roberts (2012) suggest that directives have a content to be added (when successful) to a collection of propositions that represent the mutually known active projects of the interlocutors, a "To Do List" or "Plan Set".

[14]Although Haslanger considers work on concepts in her work, she does not focus on the underlying discourse mechanisms associated with language use in public social settings, or on the role of social emotions in social movements.

4.2 Harmful perlocutionary effects or constitutive norm erosions?

The main question this paper addresses is *how can we assess the legitimacy of ameliorative purposes and projects?* The literature on conceptual ameliorations has not devoted sufficient attention to the possibility of meaning perversions. However, conceptual or meaning revisions can be either ameliorative or perversive. Ameliorations are improvements on existing meanings. Perversions, in contrast and by analogy, are corruptions of existing meanings.

This is still fairly imprecise, but it can be understood in two ways. First, perversions can be understood in terms of their harmful consequences. These could include many of the effects that Klemperer and Gessen mention: an impoverished experience, a destitute language, the loss of a shared reality, the loss of individual autonomy to a language that 'thinks for us, and dictates our feelings"', as well as actual discrimination or oppression. The political dangers of these effects should not be neglected or minimized. These are often the effects autocrats intend, since they diminish a population's capacity to resist the autocrat's control over social and political reality.

The most straightforward way to characterize meaning perversions, on the basis of Klemperer's and Gessen's testimonies, is in terms of their harmful consequences:

Causal meaning perversions are attempts to hijack language (e.g. of justice, politics, social roles, or moral or epistemic virtues) in a way that makes people worse off. In other words, meaning perversions are revisionary or engineering projects that cause harm.

In a very literal sense, if a meaning revision produces harmful consequences it is not an *amelioration*. We may be concerned, however, that the consequentialist notion of perversion cannot guide us in distinguishing between the merely feasible projects (in Podosky (2018)'s sense) and the morally legitimate projects at which we should aim. If the causal consequentialist sense of meaning perversion can't guide us in distinguishing morally legitimate from morally illegitimate projects, then it can't help us decide whether to engage in a meaning revision. We need a more perspicuous way of distinguishing ameliorations from perversions, one that goes beyond a focus on harmful *consequences*. The reason is, fundamentally, epistemic. Shelly Kagan spells out the reason why:

> Perhaps the most common objection to consequentialism is this: it is impossible to know the future. This means that you will never be absolutely certain as to what all the consequences of your act will be. An act that looks like it will lead to the best results overall may turn out badly, since things often don't turn out the way you think they will. Something extremely unlikely may happen, and an act that

> was overwhelmingly likely to lead to good results might – for reasons beyond your control – produce disaster. Or there may be long term bad effects from your act, side effects that were unforeseen and indeed unforeseeable. In fact lacking a crystal ball, how could you possibly tell what all the effects of your act will be? So how can we tell which act will lead to the best results overall – counting all the results? This seems to mean that consequentialism will be unusable as a moral guide to action. All the evidence available at the time of acting may have pointed to the conclusion that a given act was the right act to perform – and yet it may still turn out that what you did had horrible results, and so in fact was morally wrong. (Kagan, 1998, 64)

Kagan's argument raises a problem for conceptual engineering projects generally.[15] It is a problem for efforts to discriminate *ameliorations* from *perversions*, and presses us to find a more useful way of identifying which engineering concepts we *should* be engaged in, morally and politically, besides what we now believe or feel will bring about the best overall consequences. For all we know, we may be unable to know what we should do. We can hope that we can at least avoid pursuing those actions that predictably bring about harm. Yet, since we cannot easily foresee all the harm our actions can bring, it would be good to have a more perspicuous understanding of what meaning perversions can be, one that can more easily augment our ability to detect, and avoid or denounce, harmful revisionary projects.

A second sense of meaning perversions, I suggest, is *constitutive*. The cases that Klemperer and Gessen focus on, together with the theoretical resources introduced in the previous subsection, give us an indication about how to proceed. I will submit below a definition of meaning perversion that modifies Stanley (2015)'s notion of undermining propaganda, which he defines thus:

> *Undermining Propaganda*: A contribution to public discourse that is presented as an embodiment of certain ideals yet is of a kind that tends to erode those very ideals. (Stanley, 2015, 52-3).

Now, I don't simply use Stanley's definition of undermining propaganda for two reasons. First, propaganda as he defines it includes uses of code words or dogwhistles, it will also include what Saul (2017) calls racial figleaves, bald-faced lies, and other phenomena that should be distinguished from meaning perversions.[16] Second, I would like to contend that meaning perversions can occur outside the remit of propagandistic discourse, and display features that distinguish them from, e.g., code words. Hence, not all propaganda is a meaning perversion, and not all meaning perversions are propaganda.

[15] This argument is made more precise in Lenman (2000).

[16] Khoo (2017) convincingly argues that code words don't encode semantically racist content. Although this is not the place to argue for this, I think that the normative or evaluative connotation of code words is arrived at through a conversational implicature.

In my proposed definition of the constitutive sense of meaning perversions, a crucial idea is that of an *undermining norm enforcement*. Normally, a correct use of word that expressively presupposes a norm or a value contributes to reinforce that value or norm that is taken for granted as common ground (say, thanking a child for *being polite* helps to reinforce polite interactions). But meaning perversions subvert this process because we are calling things what they are not.

Constit. meaning perversions A speaker S perverts the meaning of a word w just in case S's use of w is presented as an enforcement or application of norms or values that w expressively presupposes, which erodes those very same norms or values by being misapplied to an unsuitable referent.

A use of a word is a perversion when it is false that the word applies to what the speaker intends to refer to, and the use of the word nonetheless has the *illocutionary effect* it constitutively has – it expresses a conative state to the effect that what the speaker is referring realizes a certain value or norm. One who accepts the utterance making this expressive presupposition is one who comes to take for granted that what is referred realizes that value. Since this is accepted under what I called a *sentimental mode*, one comes to take for granted the permissibility of the relevant attitudes, and possibly also to share them. These attitudes are themselves motivating and justifying reasons for jointly intentional action. The use is perversive in that what is referred does not actually realize the presupposed value. And hence the motivated actions are not the appropriate actions to take towards the presumed referrent.

What goes on with meaning perversions, like Putin's 'the dictatorship of the rule of law', the Nazi 'fanatic hero', Medvedev's 'managed democracy', Soviet elections as 'free expressions of citizen will', is that they take advantage of the different levels of content these phrases encode. There are, on the one hand, legitimate referents of words like 'democracy', 'free', 'elections'. These would be picked by the set of concrete features that correspond to the first dimension of dual-character concepts. There is, on the other hand, the expressive normative or evaluative presupposition that these phrases express. This would correspond to the set of abstract values those concrete features would realize. Thus, a sincere literal use of 'democracy' made by a competent speaker denotes any form of political organization or government that displays some minimal features (allowing for more and less fitting cases – from full to flawed democracies). To accept that some state is a democracy licenses certain presuppositions, not just about features of that state's form of government, but also about how that state realizes certain desirable normative values. To take for granted that a form of government is democratic is to take for granted that it is *desirable as good*. Mutatis mutandis, the same can

be said for 'freedom' or 'elections'.

Besides these illocutionary effects, there are additional probable harms that are perlocutionary effects of the use of the word w. The constitutive sense of meaning perversions and their perlocutionary effects are related. One sense in which we can pinpoint the nature of meaning perversions is that they have the harmful perlocutionary effects they have precisely because they contribute undermine presupposed norms or values. But the two senses are not equivalent and may not be co-extensional. Some revisions that have harmful consequences will not be constitutively perversive.

Take 'free elections'. Free elections are good things, they are essential for democracies, a recognition of the citizens' sovereignty through their representation in the institutions of their countries. The expression has a positive expressive connotation. That positive aspect of the meaning of 'free election' can be taken for granted, at least for a good amount of time, to manipulate people. By describing the ritual practice of the so-called elections in the Soviet Union (where people were forced to vote, although there was only one pre-filled ballot, and which as a result were neither free nor a real exercise of citizenship) the Soviet regime was perverting the meaning of 'free elections'. The ritual that was called a 'free election' did not display any of the concrete features that democracies must minimally exemplify. And hence, that ritual did not actually realize the positive normative value of free elections.

The normative presupposition expressed by 'free election', which is taken for granted as part of the motivational conversational set, pragmatically contradicts with the actual application of the phrase to something that does not meet the minimal constraints for being a free election. By doing this, the Soviet regime was eroding the positive value of 'free election', while normalizing the new undemocratic practice. People lived under a pragmatic contradiction between the official normative ideal – democracies are valuable and participating in them is desirable –, and the reality they were forced to inhabit. And people who are deprived of the means to appropriately describe the situation they live in are people that are deprived of the means to appropriately address it, and are more easily controlled by authoritarian regimes, again as both Klemperer and Gessen testify. These are some of the harmful perlocutionary effects of meaning perversions.

The catholic church historically pulled a most amazing meaning perversion. The word 'catholic' means all-embracing, all inclusive. The catholic church was presumed to be the universal church. And yet, for a good part of history, if you didn't believe the doctrine of the Church, or didn't do what it told you, you'd be very literally *excluded*. First, you'd be

excommunicated as a *heretic*, but you could also be *burnt at the stake*.

Meaning perversions contrast with *code words*. Code words induce the acceptance into the motivational common ground of evaluative dispositions, plans, and norms, that conflict with pre-accepted dispositions/plans/norms that are part of a shared conversational score. But the expression of those values or norms is not encoded (not even as a presupposition) in the meaning of a code word. This is exemplified in uses of code words lacking a negative racial connotation, for instance "we are doing a census of the people living in the inner city to determine the investment in new and much-needed pre-schools and centers for primary medical attention". Moreover, even those uses that do entail a racial connotation denote their *proper referent* in the world, e.g. 'inner-city' refers to actual urban areas. That's why plausible deniability is possible – one can always point out that what one said is factually true, and since the racial connotation is not part of encoded meaning, there is no contradiction in that denial.

In contrast, the normative and evaluative connotation of a meaning perversion is encoded in the word's meaning, and thus it is automatically taken for granted in the conversational record. Hence, speech that perverts meaning is easily accepted in a context, since it aligns with what is already accepted as part of the shared motivational set. This explains Klemperer (2000)'s description: "as soon as this concept (heroism) was touched upon, everything became blurred in the fog of Nazism". However, meaning perversions are not used to denote their proper referents in the world. As a result, the evaluative connotation conflicts with the use the word to talk about an *improper referent* that doesn't fit the value it is presented as realizing.

As a result of these differences, the argumentative strategies of speakers who use code words and those who use meaning perversions differ. Code words allow *plausible deniability*: "I wasn't saying anything about race! I was just talking about criminality in certain urban areas". Meaning perversions, in contrast, allow for *rhetorical norm-enforcement questions*: "How can you be against our freedom?", "How can you oppose democracy?", "How can you believe what an enemy of the people says?" Interlocutors are, naturally, expected to reject that they are against freedom, or that they oppose democracy, or that they believe their enemies. They are also now pressed to accept that an improper referent has the concrete features that realize the abstract values that are already taken for granted.

5 Closing remarks

How do the two senses of meaning perversion drawn in the previous section help us demarcate the set of morally legitimate meaning engineering projects? Revisionary projects that are *permissible* are the set of meaning revisions that are not perversions. That means that they are meaning revisions that (a) don't have harmful consequences, and that (b) do not misapply a word to something unfitting the abstract values that the use of the word presupposes.

It may be hard to know in advance if a given revision will have harmful consequences, although we can try to fend them off for instance by combining, as I suggested, reflective equilibrium methods guided by defensive pessimism and honest humility. This means essentially that we should expect bad consequences to occur (and try to foresee them to the best of our ability) and be prepared to revise our expectations and plans if necessary.

When a revision amounts to a perversion in the constitutive sense, it can be especially hard to resist. It is hard to give a reply to propagandistic calls for respecting the will of the people, or for holding free elections, or for taking back control. Yet, in the mouth of many demagogues, these are meaning perversions: uses of 'the people' that exclude most of the people, of 'free elections' that are neither free nor an exercise in autonomous individual choice, of 'take back control' that give away control, and of 'catholic' that are not all embracing and inclusive.

How do we resist meaning perversions, or spot them? Any direct criticism invites replies like "how can you be against my freedom?", "how can you be against the people?" An interlocutor is left speechless, since in normal circumstances a normal reply would be obviously "no! I'm not against the people, and I'm not against free elections". Those rhetorical questions are effective to advance a meaning perversion because they seem to reinforce shared norms, while eroding them. We, as theorists possibly interested in advancing a revisionist project, can easily be unaware that we are putting forward a perversion instead of an amelioration. If, as Maynard and Benesch (2016) put it, we are moved by "deep and unreflected feelings that something feels 'good' or 'bad"' which induce "positive moral self-appraisal", we may resist taking the extra step required to disentangle our conviction in the intended good results and the appraisal that the misapplication will erode the very same values or normative principles we believe we are promoting.

This is a small step forward in delimitating the scope of the answer to the question *how can we assess the legitimacy of ameliorative projects.* We can complement the discussion about why it is problematic to engineer race-talk, for instance. Recall that race-talk has had, and continues to have, hugely harmful consequences. Ordinary race concepts are presumed to

track a set of concrete biological features that identify a natural class, which turns out to be empty. The values that are supposed to be realized by those concrete features present people as *less*, or *more*, deserving of consideration, social standing, or respect as persons by being presumed to exhibit certain biological features. We can try to ameliorate what 'race' refers – for instance, that it refers to ancestor-descendant sequences of breeding populations that share a common origin. But the normative presuppositions that come with race-talk are not amenable to amelioration by *fiat*, particularly when there have been billions killed or enslaved on account of the negative values they were presumed to realize. Decisions to do away with talk of *Rasse* or *raça* acknowledge this heavy burden. Moreover, there are alternative ways to address resilient racism, and to pursue useful lines of inquiry in biology, anthropology, or sociology.

In this paper, I argued for the importance of adding the question *How can we assess the legitimacy of ameliorative projects?* to the list of other questions pursued by theories about concepts of social roles and categories: What is the nature of the social things we talk about? How do we think about the social world, and how does our thinking relate to the nature of social reality? What is the point of having the concept in question? What concept (if any) would better serve our social or political goals? And who are we? I have also argued that our focus on normative constraints on conceptual engineering should include a reflection on the moral and political legitimacy of the projects pursued. I was motivated, in this regard, by lessons from deeply problematic historical cases. By trying to offer characterizations of meaning or conceptual perversions, the contrary of ameliorations, I hope to have offered a way of demarcating what a legitimate conceptual amelioration *cannot be*.

References

Andreasen, R. O. (1998). A new perspective on the race debate. *British Journal for the Philosophy of Science 49*(2), 199–225.

Andreasen, R. O. (2000). Race: Biological reality or social construct? *Philosophy of Science 67*(3), 666.

Andreasen, R. O. (2004). The cladistic race concept: A defense. *Biology and Philosophy 19*(3), 425–442.

Andreasen, R. O. (2005). The meaning of 'race'. *Journal of Philosophy 102*(2), 94–106.

Appiah, K. A. (2006). How to decide if races exist. *Proceedings of the Aristotelian Society 106*(3), 363–380.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, C. (2017). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press USA.

Björnsson, G. and K. Hess (2017). Corporate crocodile tears? On the reactive attitudes of corporate agents. *Philosophy and Phenomenological Research 94*(2), 273–298.

Cepollaro, B. and I. Stojanovic (2016). Hybrid evaluatives: In defense of a presuppositional account. *Grazer Philosophische Studien 93*, 458–488.

Charlow, N. (2016). Decision theory: Yes! truth conditions: No! In N. C. M. Chrisman (Ed.), *Deontic Modality*. Oxford University Press.

Couto, A. (2019). Reactive attitudes, forgiveness, and the second-person standpoint. *forthcoming in Ethical Theory and Moral Practice*, 1–15.

Diaz-Leon, E. (2015). What is social construction? *European Journal of Philosophy 23*(4), 1137–1152.

Figes, O. (2002). *The Whisperers: Private Life in Stalin's Russia*. London: Penguin Books.

Gessen, M. (2017). The autocrat's language. *The New York Review of Books May 13, 2017*.

Glasgow, J. (2003). On the new biology of race. *Journal of Philosophy 100*(9), 456–474.

Glasgow, J. (2019). Conceptual revolutions. In T. Marques and Åsa Wikforss (Eds.), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variability*. Oxford: Oxford University Press. Forthcoming.

Goldman, D. (2014). Modification of the reactive attitudes. *Pacific Philosophical Quarterly 95*(1), 1–22.

Haslanger, S. (2003). Social construction: The "debunking" project. In F. Schmitt (Ed.), *Socializing Metaphysics*, pp. 301–325. Oxford: Rowman and Littlefield.

Haslanger, S. (2006). What good are our intuitions? Philosophical analysis and social kinds. *Aristotelian Society Supplementary Volume 80*(1), 89–118.

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

Haslanger, S. (2018). Nonideal, critical, ameliorative, realist social ontology: Say what? Keynote lecture at *Social Ontology 2018* conference in Boston.

Haslanger, S. (2019). Going on, not in the same way. In D. P. Alexi Burgess, Herman Cappelen (Ed.), *Conceptual Ethics and Conceptual Engineering*, pp. forthcoming. Oxford: Oxford University Press.

Jeshion, R. (2016). Slur creation, bigotry formation: the power of expressivism. *Phenomenology and Mind 11*.

Kagan, S. (1998). *Normative Ethics*. Boulder: Westview Press.

Khoo, J. (2017). Code words in political discourse. *Philosophical Topics 45*(2), 33–64.

Kitcher, P. (2007). Does 'race' have a future? *Philosophy and Public Affairs 35*(4), 293–317.

Klemperer, V. (1957/2000). *The Language of the Third Reich – LTI Lingua Tertii Imperii, a Philologist's Notebook*. London: Continuum.

Knobe, J. and S. Prasada (2011). Dual character concepts. In *Proceedings of the* 33^{rd} *Annual Conference of the Cognitive Science Society*. Boston, MA: Cognitive Science Society.

Knobe, J., S. Prasada, and G. Newman (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition 127*(2), 242–257.

Kutz, C. (2000). Acting together. *Philosophy and Phenomenological Research 61*(1), 1–31.

Langton, R. (2012). Beyond belief: Pragmatics in hate speech and pornography. In M. K. McGowan and I. Maitra (Eds.), *Speech and Harm: Controversies Over Free Speech*, pp. 72. Oxford University Press.

Lawford-Smith, H. (2013). Understanding political feasibility. *Journal of Political Philosophy 21*(3), 243–259.

Lenman, J. (2000). Consequentialism and cluelessness. *Philosophy and Public Affairs 29*(4), 342–370.

Lippert-Rasmussen, K. (2013). *Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination*. Oxford University Press.

Ludwig, D. (2018). How race travels. relating local and global ontologies of race. *Philosophical Studies*, 1–22.

Machery, E., R. Mallon, S. Nichols, and S. Stich (2009). Against arguments from reference. *Philosophy and Phenomenological Research 79*(2), 332–356.

Manne, K. (2017). *Down Girl: The Logic of Misogyny*. Oxford University Press.

Marques, T. and M. García-Carpintero (2019). Really expressive presuppositions and how to block them. *Grazer Philosophischen Studien*. forthcoming in special issue on *Non-derogatory issues of slurs* edited by Bianca Cepollaro and Dan Zeman.

Maynard, J. and S. Benesch (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention: An International Journal 9*(3), 70–95.

Müller, K. and C. Schwarz (2018). Fanning the flames of hate: Social media and hate crime. Available at SSRN: https://ssrn.com/abstract=3082972 or http://dx.doi.org/10.2139/ssrn.3082972.

Norem, J. (2001). Defensive pessimism, optimism, and pessimism. In E. Chang (Ed.), *Optimism & pessimism: Implication for theory, research, and practice*, pp. 77–100. Washington DC: American Psychological Association.

Podosky, P. C. (f2018). Ideology and normativity: Constraints on conceptual engineering. *Inquiry: An Interdisciplinary Journal of Philosophy*, 1–15.

Portner, P. (2016). Imperatives. In M. Aloni and P. Dekke (Eds.), *Cambridge Handbook of Semantics*, pp. 593–628. Cambridge University Press.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics 49*(5), 1–69.

Salmela, M. and N. Michiru (2016). Collective emotions and joint action. *Journal of Social Ontology 2*(1), 33–57.

Salmela, M. and C. von Scheve (2017). Emotional roots of right-wing political populism. *Social Science Information 56*(4), 567–595.

Saul, J. (2017). Racial figleaves, the shifting boundaries of the permissible, and the rise of Donald Trump. *Philosophical Topics 45*(2), n/a.

Simion, M. (2017). The *should* in conceptual engineering. *Inquiry: An Interdisciplinary Journal of Philosophy 61*(8), 914–928.

Snyder, T. (2017a). The deliberate starvation of millions in Ukraine. *The Washington Post November 3*. Review of Anne Applebaum's *Great Famine - Stalin's War on Ukraine*.

Snyder, T. (2017b). *On Tyranny: 20 lessons for the 20th century*. New York: Duggan Books.

Stalnaker, R. (1978). Assertion. *Syntax and Semantics (New York Academic Press) 9*, 315–332.

Stamp, N. (2018). Women with heart diseases are dismissed and its killing them. *The Guardian*. https://www.theguardian.com/commentisfree/2018/jun/14/women-with-heart-diseases-are-dismissed-and-its-killing-them.

Stanley, J. (2015). *How Propaganda Works*. New Jersey: Princeton University Press.

Stemplowska, Z. and A. Swift (2012). Ideal and non-ideal theory. In D. Estlund (Ed.), *The Oxford Handbook of Political Philosophy*, pp. 373–389. Oxford: Oxford University Press.

Strawson, P. (1962/2008). Freedom and resentment. *Proceedings of the British Academy 48*, 1–25. reprinted in *Freedom and Resentment and Other Essays*, 2nd edition, pp. 1-28, New York: Routledge.

Tarski, A. (1943). The semantic conception of truth: And the foundations of semantics. *Philosophy and Phenomenological Research 4*(3), 341–376.

Tirrell, L. (2012). Genocidal language games. In I. Maitra and M. K. McGowan (Eds.), *Speech and Harm: Controversies Over Free Speech*, pp. 174–221. Oxford University Press.

Wilkins, D. B., K. A. Appiah, and A. Gutmann (1998). *Color Conscious: The Political Morality of Race*. Princeton University Press.

Williams, B. (1979). Internal and external reasons. In R. Harrison (Ed.), *Rational Action*, pp. 101–113. Cambridge University Press.

Yanagizawa-Drott, D. (2014). Propaganda and conflict: Theory and evidence from the Rwandan genocide. *The Quarterly Journal of Economics 129*(4), 1947–1994.