Epistemic paternalism via conceptual engineering

Eve Kitsik (University of Cologne)

Forthcoming in *Journal of the American Philosophical Association*

Abstract. The paper targets conceptual engineers who aim to improve other people's patterns of inference and attention by shaping their concepts. Such conceptual engineers sometimes engage in a form of epistemic paternalism that I call "paternalistic cognitive engineering": instead of explicitly persuading, informing and educating others, the engineers non-consultatively rely on assumptions about the target agents' cognitive systems to improve their belief-forming. The target agents could reasonably regard such benevolent exercises of control as violating their sovereignty over their own belief-formation. This is a *pro tanto* reason against such engineering. In addition to the relevant projects of conceptual engineering, paternalistic cognitive engineering plausibly includes certain kinds of nudging and evidence suppression. The paper distinguishes the sovereignty-based concern from other ethical worries about conceptual engineering and discusses how one might justify the relevant conceptual engineering projects despite the sovereignty-based reason against them.

Keywords: epistemic paternalism, conceptual engineering, nudging, evidence suppression, epistemic autonomy

1. Introduction

"Epistemic paternalism" means non-consultatively interfering with someone's inquiry or belief-formation for their own epistemic good. Examples include withholding true and relevant but potentially misleading evidence from jurors (Goldman 1991), demoting fake news on social media (Castro, Pham, and Rubel 2020), and nudging someone toward true beliefs by having members of their political party present the evidence (McKenna 2020). "Conceptual engineering" means evaluating and improving (or at least trying to improve) concepts. Examples include repairing inconsistent concepts, such as (allegedly) the concept of truth (Scharp 2013); engineering gender and race concepts to serve social progress (Haslanger 2000); and engineering the term "belief", as used in philosophy, to designate the most important kind in the vicinity (Schwitzgebel 2021). This paper argues that conceptual engineering can involve a problematic form of epistemic paternalism.

I limit the discussion to conceptual engineers who aim to improve the concept users' patterns of inference and attention. Concepts, on such an approach, are psychological entities in individual minds and play an important role in cognition (see Machery 2017, ch. 7; Fischer 2020; Isaac 2020; Isaac 2021a; Isaac 2021b; Koch 2021; Machery 2021). The relevant kind of conceptual engineer recognizes the power of concepts over our thought and deliberation, and wants to use that power for the good, including the concept users' own epistemic good, broadly construed. Such conceptual engineering contrasts with the variety that aims to change the meanings of words in a language, where these meanings may be determined by inscrutable factors external to the concept users' minds (e.g., Cappelen 2018). Although I only raise worries for conceptual engineering that is concerned with concepts as psychological entities, the idea is not that this kind of conceptual engineering is particularly problematic. On the contrary, I take it to be the best motivated among the approaches to conceptual engineering on

offer. That makes it even more important to discuss whether, when and how these otherwise compelling conceptual engineering projects engage in problematic epistemic paternalism.

The paper is structured as follows. In section 2, I stipulatively adopt Ahlstrom-Vij's (2013) wide definition of "epistemic paternalism". Within this wide category, I identify a set of *pro tanto* problematic practices that I label "paternalistic cognitive engineering". These practices rely on assumptions about the target agents' cognitive system to non-consultatively shape their belief-formation for their own epistemic good, in a way that these agents could reasonably regard as violating their sovereignty over their domain of belief-formation. In section 3, I argue that some conceptual engineering projects involve such paternalistic cognitive engineering. I provide examples of the relevant projects and distinguish the sovereignty-based concern raised here from other ethical concerns about conceptual engineering. In section 4, I discuss how one could justify such conceptual engineering projects despite the sovereignty-based reason against them, either by appealing to the target agents' epistemic improvement alone or by appealing to the epistemic improvement's broader social impact. Section 5 offers concluding remarks.

2. From epistemic paternalism to paternalistic cognitive engineering

This section aims to delineate a *pro tanto* problematic variety of epistemic paternalism that the relevant conceptual engineering projects plausibly belong to. Merely showing that these projects involve a form of "epistemic paternalism" in *some* sense of that term would not raise a worry for the projects. It is possible, after all, to define "epistemic paternalism" in a way that does not isolate a set of (even *pro tanto*) problematic practices. Indeed, Ahlstrom-Vij's influential account of epistemic paternalism arguably defines the term in such a way. According to Ahlstrom-Vij (2013: 38–61), the necessary and sufficient conditions for an epistemically paternalistic practice are the following.

(1) Interference: The agent's ability to go about their inquiry (to access, collect, and evaluate information) in whatever way they see fit is interfered with.

(2) Non-Consultation: The agent is not consulted, meaning that either they are not asked about the interference, or they are asked but then not listened to.

(3) Improvement: The aim is to improve the agent's epistemic situation (which may be instrumental for further non-epistemic goals).

For example, this definition covers the central case of epistemic paternalism discussed by Goldman (1991): withholding true and relevant but nevertheless misleading evidence (such as information about the defendant's past crimes) from jurors. Such evidence suppression is expected to improve the jurors' cognitive performance, because people tend to attribute too much weight to the kinds of evidence in question. The practice satisfies all three conditions: it interferes with the jurors' ability to inquire in whatever way they see fit; the jurors are not consulted; and the aim is to help them reach the true verdict—their own epistemic good—which is instrumental for further aims, such as justice being served.

Defined this way, epistemic paternalism arguably includes many entirely unproblematic practices. As Medvecky (2020) points out, everyday testimony seems to fit the bill. We choose what to say and what to leave out, so that our audience gets the knowledge that fits their needs and interests. For example, when telling someone the time, we often round up to the nearest minute or more, thus interfering with the hearer's inquiry into the current time with their epistemic well-being in mind and without consulting them. Further, even writing a non-fiction book might count as non-consultatively interfering with its readers' inquiry for their own epistemic good (Jackson 2021: 135).

Ahlstrom-Vij meant to capture the problematic aspect of epistemic paternalism in the non-consultation condition, which should explain why there is "something arrogant about paternalistic interference; a distinct kind of disregard for the wants, preferences or opinions of

those interfered with" (Ahlstrom-Vij 2013: 43). But as the everyday testimony and book-writing examples show, non-consultation as such is not a problem if the interference is mild enough. We need not obtain consent every time we help people out in small ways.

Medvecky takes such examples (in particular, about everyday testimony and science communication) to show that epistemically paternalistic practices are generally just fine. Jackson (2021), by contrast, takes cases of innocuous (supposed) epistemic paternalism to show the inadequacy of the definition of epistemic paternalism that includes these cases. For my purposes, the discussion on how to define "epistemic paternalism" can be largely left aside. Merely as a convenient stipulation (which will not matter for the argument to come), I reserve the term for (roughly) the broad phenomenon identified by Ahlstrom-Vij. I will be concerned with a subset of *pro tanto* problematic practices within such broadly construed epistemic paternalism—a subset that, I will argue, includes the relevant projects of conceptual engineering. I call this subset "paternalistic cognitive engineering". The practices in this subset have the following features.

(1) Objectionable cognitive engineering: The practice involves *cognitive engineering*, that is, deliberately and reflectively shaping an agent's belief-formation in light of assumptions about their cognitive system. Further, the cognitive engineering is *objectionable* in the sense that the agent could reasonably regard it as a violation of their sovereignty over their belief-formation (assuming that they are not consulted).

(2) Non-Consultation: The agent is not consulted (in the same sense as in Ahlstrom-Vij's account).

(3) Improvement: The aim is to improve the agent's epistemic situation. As in Ahlstrom-Vij's account, this may be instrumental for further non-epistemic purposes. (But note that I understand "epistemic" more broadly than Ahlstrom-Vij; in particular, attending to important matters counts as "epistemic" improvement here.)

The non-consultation and improvement conditions are the same as in Ahlstrom-Vij's account, then. The objectionable cognitive engineering condition needs unpacking. I use "cognitive engineering" similarly to how Hausman and Welch (2010) use "shaping" in their critical discussion of nudging. "Nudging", a notion introduced and popularized by Thaler and Sunstein (2008), means making subtle changes to the agents' environment to non-coercively influence their choices, often for their own good; for example, placing healthy food at eye level or setting relatively high default values of retirement savings. Hausman and Welch point out that some nudges involve "shaping" in the sense of "the use of flaws in human decision-making to get individuals to choose one alternative rather than another"; they contrast such shaping with rational persuasion (Hausman and Welch 2010: 128). However, "cognitive engineering", as I use the term, need not use *flaws* in human decision-making: it can rely on other assumptions about human cognition or specifically the target agents' cognition. Another difference is that I emphasize the shaping of *belief-formation* rather than choices. But there might not be a principled distinction here: good choices may be construed as true beliefs about what is to be done.

Cognitive engineering is reflective and deliberate. For example, we do not engage in cognitive engineering when we unreflectively round off the time for the hearer's epistemic good—though we may be unconsciously relying on knowledge about human cognitive limitations and the hearer's needs. This is not to say that if we were to consciously consider the hearer's limitations and needs, and deliberately round off the time in this light, this would be paternalistic cognitive engineering. Whether or not it makes sense to count this as "cognitive engineering", the "objectionability" part of the first condition is unsatisfied: it would not be reasonable for the hearer to regard our act as a sovereignty violation.

Let us look at the "objectionability" part of the first condition, then. It appeals to the familiar idea of personal sovereignty, "the idea of having a domain or territory in which the self is

sovereign" (Feinberg 1983: 452). As Feinberg further explains this, "to say that I am sovereign over my bodily territory is to say that I, and I alone, decide (so long as I am capable of deciding) what goes on there"; and what falls within one's sovereign domain "cannot be treated in certain ways without one's consent" (ibid.: 453). The violations that we are here concerned with, however, are not essentially interferences with the body or its surrounding space but rather with the domain of belief-formation. Mill, suitably for our purposes, emphasizes sovereignty over both the body and the mind: "over himself, over his own body and mind, the individual is sovereign" (Mill 1999/1859: 52).

The idea, then, is that we all (or at least the reasonably competent adult humans among us) have our territory of belief-formation that we are entitled to govern ourselves and that should not be treated in certain ways without our consent. In paternalistic cognitive engineering, the engineers non-consultatively exercise control over the target agents' domain of belief-formation in a way that these agents could reasonably regard as intruding into that territory and thus as violating their sovereignty over their own intellectual domain, their own mind. That is a *pro tanto* reason against such engineering.

I have purposefully used a cautious formulation, appealing to how the target agents could reasonably assess the practice, instead of simply asserting that the relevant practices *are* sovereignty-violating. The reason for the cautious formulation is that people differ widely in what practices they would regard as intruding into their territory; and it is plausible that there is a range of reasonable attitudes here, rather than a clear division between objectively intrusive and non-intrusive practices. For example, some people regard certain benevolent nudges (such as healthy food placed at eye level) as disturbingly controlling their choices, while others do not mind or even appreciate the nudges. I want to acknowledge a range of reasonable attitudes, which leads me to refrain from claims about objective sovereignty-violation.

At the same time, we should not classify practices as *pro tanto* problematic just based on the recipients' actual or hypothetical judgment: people sometimes unreasonably feel intruded. For example, people can feel intruded when we politely challenge their political convictions with good arguments; and this is not a sovereignty-based reason against such polite rational persuasion. So, "objectionability" is not a matter of whether the target agents *would* or *could* assess the practice as sovereignty-violating; it depends on whether regarding the practice as sovereignty-violating falls within the range of reasonable attitudes for the target agents (were they to find out afterwards, for example, or were they to discover that such a practice is being planned).

Further, I assume that it is reasonable for the target agents to regard a practice as sovereignty-violating when the paternalistic actors exercise significant control over the target agents' belief-formation by cognitive engineering, while it would in principle be possible to explicitly persuade, inform or educate the agents instead. This to some extent draws on and generalizes Hausman and Welch's criticism of nudges of the "shaping" variety. According to Hausman and Welch (2010), such nudges infringe on autonomy in the sense of "the control an individual has over her own evaluation, deliberation and choice" (128). Granted, some people might not care much about controlling their own intellectual domain, and this attitude might be reasonable. However, I assume (not outrageously, I hope) that it is also reasonable to wish that others did not exercise significant control by cognitive engineering—especially when it is possible to instead engage in explicit persuading, informing and educating, thus giving the target agents more control.

For example, suppose that a doctor consciously uses a framing effect to nudge the patient to decide in favor of an operation, for the patient's own good, and says: "The chances of survival are 99%" (rather than: "There is 1% chance of dying"). They could have instead stated their recommendation explicitly, along with the statistical information and the reasoning behind the

recommendation, thus giving the patient more control over their decision. Since this alternative approach is available, the patient (were they to find out about the intervention) could reasonably think that the doctor is exercising too much control over the patient's intellectual domain.

For another example, Riley (2017: 610) contrasts two ways of combatting low electoral turnouts: (a) implementing a program of beneficent nudges and (b) setting aside three days before the election for public discussion and debate, to raise the community's level of political engagement. Here as well, exercising control by nudging seems problematic because there is a less controlling method available for pursuing the target agents' own good; and those agents could reasonably wish that this method were used instead of cognitive engineering.

An important clarification is that explicit rational persuasion, informing and educating can be "available" even when they are more difficult, costly and time consuming to implement than cognitive engineering, or not quite as effective. For example, consider the practice of evidence suppression, which relies on assumptions about the target agents' cognition (the relevant biases) to shape belief-forming, and is thus a variety of cognitive engineering. An alternative approach available here is educating the target agents about their relevant biases and suggesting strategies to counteract bias. That alternative approach would give these agents more control over their intellectual domain, but it would also be much more costly and time-consuming and might ultimately leave the agents worse off in terms of reliable belief-formation, compared to external evidence curation. Nevertheless, the educating approach is plausibly "available" enough for the target agents to reasonably regard non-consultative evidence suppression as a sovereignty violation. (However, reasonably taking a practice to be sovereignty-violating is not the same as reasonably taking it to be ultimately unjustified. Perhaps it is reasonable for the jurors, for example, to regard evidence suppression as sovereignty-violating—so, there is a *pro tanto* reason against evidence suppression—while it is unreasonable for the jurors to regard the

practice as ultimately unjustified, given the reasons in favor of the practice. I will return to the issue of ultimate justification in section 4.)

Since I contrast paternalistic cognitive engineering with explicit rational persuasion, informing, and educating, it is also important to emphasize that the problem with paternalistic cognitive engineering is *not* that it bypasses rationality. "Bypassing rationality" means influencing choice, belief and behavior without providing reasons. Levy (2019; 2022) argues that nudges often provide *implicit* reasons and therefore do not bypass rationality. For example, according to Levy, when a doctor uses a framing effect to make salient the high probability of surviving the operation, the doctor is implicitly recommending the operation and is interpreted as such by the patient. Further, the doctor's opinion is a reason—moreover, a good reason—for the patient to decide in favor of the operation. Even if Levy is right that nudges typically provide reasons,[1] it could still be reasonable to regard nudging as sovereignty-violating. For example, when the patient finds out that the doctor reflectively chose this framing to influence the patient's choice, the patient may well be reasonable in feeling that the doctor violated their sovereignty—not because the doctor failed to provide a reason, but because they offered a reason in a way that exercised too much control over the patient's decision. The doctor could have stated their recommendation explicitly, perhaps along with the reasoning behind it, giving the patient more control. So, the problem with paternalistic cognitive engineering is not bypassing rationality; the problem is exercising control over another person's intellectual domain in a way they could reasonably regard as sovereignty-violating.

As the above examples show, paternalistic cognitive engineering— a *pro tanto* problematic (but perhaps not ultimately unjustified) variety of epistemic paternalism—plausibly includes some cases of nudging and evidence suppression. The category does not include some other

[1] For discussion, see Grundmann 2021a, Levy 2021a, Grundmann 2021b, Levy 2021b.

typical examples of problematic epistemic paternalism, such as a doctor divulging medical information against the patient's stated will (Bullock 2018: 441)—there is no cognitive engineering here, and hence no paternalistic cognitive engineering. But this is fine: paternalistic cognitive engineering is not supposed to cover all problematic epistemic paternalism. I am only interested in the kind of (*pro tanto*) problematic epistemic paternalism that plausibly includes the relevant conceptual engineering projects. I now turn to these projects.

3. Conceptual engineering as paternalistic cognitive engineering

Let us, then, move on to reflective and deliberate attempts to shape the target agents' belief-formation *by shaping their concepts*. In order for these practices to involve paternalistic cognitive engineering, it must be reasonable for the target agents to view the practices as sovereignty-violating. As suggested above, this is reasonable if the engineers could in principle explicitly persuade, inform and educate instead of cognitive engineering, thus giving the target agents more control over their intellectual domain. Further, the practices must be non-consultative and aim at least instrumentally for the agents' epistemic improvement, broadly construed. I will argue that all the conditions are satisfied for some conceptual engineering projects and that the projects therefore involve paternalistic cognitive engineering. So, there is a sovereignty-based *pro tanto* reason against these projects. The next section will then discuss whether reasons in favor of such conceptual engineering could outweigh the sovereignty-based reason against it.

Again: concepts, in the relevant sense, are psychological entities. One may further conceive of concepts, on this approach, as bodies of information (in the long-term memory) that are drawn on quickly, automatically, and context-independently to form judgments; these bodies of information might include exemplars, prototypes, and theory-like structures (Machery 2017: 210–212). While concepts, thus understood, may be engineered for various aims, they are

sometimes engineered (at least in the first instance) for broadly epistemic aims. Let us look at a few cases of this.

Some of the relevant projects exploit concepts' capacity to underwrite inferences. As Machery (2017) puts it, "Concepts determine the inferences we are prone to draw, and our thoughts follow their tracks" (231). Sometimes these are epistemically bad tracks, and we might then want to discourage the use of such concepts. For example, Brandom (1994) suggests eliminating the slur "boche" because it underwrites unreliable inferences: "If one does not believe that the inference from German nationality to cruelty is a good one, then one must eschew the concept *Boche*" (126). Another example is Machery's (2021) case for eliminating the concept of innateness in biology. The folk concept of innateness, according to Machery, promotes unreliable inferences by bundling together the notions of typicality, fixity, and functionality. Although biologists take themselves to be using a technical concept of innateness, they still unreflectively draw the unreliable inferences stemming from the folk concept. For this reason, Machery proposes dispensing with innateness talk in biology altogether.

This is, then, the first type of project: eliminating concepts that underwrite unreliable inferences, or to put it simply, that bring to mind false ideas (such as false stereotypes) and make these ideas appear plausible.[2] The elimination of an epistemically bad concept may be achieved by explicit guidelines, social norms and sanctions, or the influence of authoritative language users, for example. The idea is that changing the community's linguistic behavior would also change individuals' habits of using concepts in their private cognition. If the concepts that support unreliable inferences cannot be completely weeded out in this manner, they might at least become less easily activated, and that is already an epistemic win. In the

[2] See Fischer (2020) for a more detailed, empirically informed explanation of how concepts underwrite inferences.

case of prejudice-evoking slurs, the epistemic wins from their elimination bring non-epistemic wins in their tow; but one may also want to eliminate concepts that support inferences to falsehoods just for the epistemic wins alone.

In other cases, the concept itself is fine and even commendable, but certain uses of it shape the inferences that the concept underwrites in problematic ways. Such uses can then be discouraged for the concept users' own epistemic good. For example, consider Fraser's discussion of rape metaphors that make bad inferences involving "rape" more cognitively accessible and socially licensed, and good inferences less accessible and less licensed. An inference is cognitively accessible when it "does not require significant cognitive labor" and socially licensed when it is generally regarded as legitimate (Fraser 2018: 735–736). For example, saying "Germany is raping Brazil!", while watching soccer, can make the inferences from "S was raped" to "S has been humiliated" and "S should have put up more of a fight" more cognitively accessible and socially licensed, while making the inference to "Serious injustice has been done to S" less accessible and licensed (ibid.: 745–746). From the epistemic perspective, such rape metaphors are problematic, then, because they shape the concept of rape in a way that disposes the concept user to make inferences from truths to falsehoods and indisposes them to make certain inferences to truths. The concept users thereby become more likely to acquire false beliefs and miss out on important truths. Again, the problematic usage can be discouraged by establishing guidelines and social norms.

The previous examples concerned concepts underwriting inferences. Conceptual engineers can also exploit the capacity of entrenched terms to make phenomena psychologically salient. For example, consider Haslanger's (2000) proposal of revising the terms "man" and "woman" to designate humans that are respectively advantaged or disadvantaged in virtue of their (real or imagined) features related to their role in reproduction. Different interpretations of this are possible; for example, Haslanger (2006) herself and Deutsch (2020a; 2020b) argue that the

proposal can be seen as one about what the terms already mean. However, another possible interpretation is that this conceptual change would facilitate a desirable change in patterns of attention: it would make the many forms of gender inequality more salient, thereby motivating progressive action. This interpretation is supported by Haslanger's (2000) remark that she is "asking us to understand ourselves and those around us as deeply molded by injustice and to draw the appropriate prescriptive inference" (48).

Clark and Chalmers's (1998) defense of an extended notion of belief (which includes externally stored easily accessible information) can be understood similarly. (Here as well, it is unimportant whether the project I am describing exactly fits the authors' intentions or is otherwise the "best" interpretation, as long as this is a somewhat compelling project in the vicinity.[3]) When we associate the entrenched term "belief" with the more natural kind in the relevant contexts, the more natural kind becomes salient for us in these contexts, facilitating effective explanation and prediction.

Schwitzgebel (2021) explicitly makes this sort of case for adopting his preferred pragmatic concept of belief within philosophy. The pragmatic concept requires, in addition to intellectual endorsement, also the relevant phenomenological and behavioral dispositions. According to Schwitzgebel, beliefs in that pragmatic sense are more important for philosophers to think about than mere intellectual endorsements. So, we should use the central term "belief" to guide attention to those more important phenomena and can use the less central term "judgment" to talk about mere intellectual endorsements.

One might question whether improving patterns of attention within philosophy or in society at large is an *epistemic* improvement, and whether these salience-shaping projects are thus candidates for paternalistic cognitive engineering, as defined above. As also noted above,

[3] In a recent paper, Chalmers (2020: 8) notes that they meant to give an account of what "belief" already means, but he does not mind if it is thought of as a conceptual engineering project instead.

however, I construe "epistemic" broadly. This contrasts with some classics on epistemic paternalism, who focus on narrowly veritistic aims (Goldman 1991; Ahlstrom-Vij 2013). In the more recent discussion, the "epistemic" in "epistemic paternalism" has often been broadly construed, in line with my approach here. Moreover, such a broad construal allows us to discuss under the convenient heading "epistemic paternalism" issues that are closely related and naturally discussed together. There also seems to be no consensus usage of "epistemic" that this approach would conflict with. Hazlett (2016) and Cohen (2016) consider this lack of consensus and boundaries to be unfortunate; but for present purposes, it comes in handy.

So, the relevant kind of conceptual engineer aims at the target agents' epistemic good, broadly construed, by shaping their inferential and attentional dispositions. The improvement condition is thus satisfied: the interference aims at the target agents' own epistemic good. Further, cognitive engineering is involved: the conceptual engineer who encourages or discourages certain concepts (or certain uses of certain concepts) deliberately and reflectively shapes belief-formation in light of assumptions about how the target agents' minds work. The assumptions concern both how concepts affect cognition generally—that they support inferences and alter salience—and how the concepts in question work (what inferences they support, how the various uses of concepts in turn shape these inferences, and what the concepts make salient). In light of such information, the conceptual engineer then shapes the target subjects' belief-formation—for example, by trying to remove undesirable (uses of) concepts from circulation or taking steps to ensure that preferred concepts are instilled.

Two questions remain. First, could the target agents reasonably regard the engineering as sovereignty-violating? This is the case, recall, if the engineers could in principle use methods like explicit rational persuasion instead, thus giving the target agents more control over their intellectual domain. And second, is the practice non-consultative?

Regarding the first issue, let us consider engineering concepts to shape inference patterns—for example, eliminating *boche* in order to eliminate or weaken the inference from "S is German" to "S is cruel"; or discouraging the relevant rape metaphors (like "Germany is raping Brazil!", about soccer) to support the inference from "S was raped" to "A great injustice has been done to S". Could the engineers engage in explicit rational persuasion instead? There are some options to consider, at least. For example, instead of removing stereotype-evoking expressions or the rape metaphors from circulation, we could undertake the more resource-demanding but arguably also more respectful project of reasoning with people about why the stereotypes are false or why rape is a great injustice. Directly addressing the mistaken ideas perpetuated by the linguistic practices (like slurs and rape metaphors) could then also lead to changes in those linguistic practices and the desired developments in concepts. Educating people about ethnic prejudice and debunking the mistaken ways of thinking about rape would plausibly leave our target agents more control over their intellectual domain, without us having to give up on the aim of epistemic improvement.

For salience-altering conceptual engineering as well, there are alternative approaches to consider. The non-consultative engineer shapes what the agents experience as salient in explaining, predicting, and interpreting, instead of rationally defending the importance of these phenomena. Such a rational defense must be possible, for the conceptual engineer to be able to justify the salience-altering program; but if the rational defense is possible, then why not offer that defense to the target agents, instead of just altering their experience by conceptual engineering? Consider Haslanger's (2000) proposal, as interpreted above. The engineer could educate the audience about how the world is thoroughly shaped by gender inequality and how to spot it in its different guises. Instead, the non-consultative engineer instils revisionary gender concepts to automatically alert us to gender inequality. The target agents, it seems, might well insist on the educating approach, to retain more control over what they notice, think about, and

feel motivated to do. Granted, the cognitive effects of the educating approach would not be exactly the same as those of the engineering approach; but it would serve the same general aim of promoting attention to the pervasiveness of gender inequality. Further, the educating approach may even include the suggestion to try looking at the world through the lens of the revisionary gender concepts—such a suggestion would not be the sort of non-consultative engineering that the educating approach contrasts with.

So, when we non-consultatively engineer other people's concepts to shape their patterns of inference and attention, they could reasonably regard this as a violation of their intellectual sovereignty, wishing that the engineers would not exercise control like this and would rather use explicit rational persuasion, informing, and educating. But are the proponents of conceptual engineering indeed envisioning these projects as non-consultative? Let us first get clearer on what "(non-)consultative" means here. The consultative conceptual engineer *offers* a conceptual revision to the target agents as a way of improving their belief-formation, along with an explanation of the aims and mechanisms involved. The agents are then free to take the advice or to refuse it. For example, the conceptual engineer could explicitly make a case to the target agents that the proposed concept would make them better at tracking the truth or that it would facilitate attending to important matters.

Indeed, this seems to be precisely what is going on in many prominent conceptual engineering proposals: the philosopher offers a way of revising a concept or suggests eliminating or introducing a concept, and the audience is free to take up the offer or refuse it. For example, Haslanger (2000) and Clark and Chalmers (1998) did not devise any clever schemes to make people adopt their preferred concepts; they just offered their arguments. But there is certainly a faction of conceptual engineers who think of the enterprise as more involved in implementing the proposals. For example, Sterken (2020) advocates linguistic transgressions—simply talking as if the word already had the desired meaning—as a

potentially effective strategy. Or consider how Pinder envisions changing speaker-meanings in larger communities:

> So she should publicise the new definition, building public momentum for its adoption. At the outset, this may involve seeking to convince a high-profile group of users to defer to her definition. Then, over time, she might broaden that group, selectively targeting individuals who are influential in different communities. She will need simple and intuitive definitions, compelling arguments and powerful examples, as well as time, influence and luck. (Pinder 2021: 157)

Arguments have a role in this vision, but so does the influence of high-profile language users. For another example, Nimtz (2021) suggests that there is no need to rationally convince all the target speakers: it might be enough to convince just a few, "given that these are the right ones" (3–4). The right ones are "those whom ordinary speakers look to in the matter" (ibid., 23). This strategy makes the intervention a non-consultative one for the concept users who are not rationally convinced.

Should conceptual engineers just refrain from engineering non-consultatively, then, and instead rationally convince all the target agents to adopt the proposed concepts? I do not rule this out completely, but complications should be noted. First, the engineers sometimes seem to have in mind large populations as their target; and indeed, it often makes sense to target large populations. For example, it is not enough to make a few friends stop thinking about rape in the mistaken ways facilitated by rape metaphors; this is a culture-level problem and needs to be addressed as such. However, it would be a tough feat to even reach such large populations with one's arguments, much less convince them all.

Further, even when we only target a smaller group of people that we can hope to reach and convince—for example, our fellow philosophers—there are complications. For shaping

inference and attention patterns, it might not be enough that those rationally convinced make a conscious effort to use terms in a certain way. As Machery (2021) and Fischer (2020) both point out, our ability to consciously control the concepts that we use is limited. Even if one decides to use a term in a revised way in a certain context, the folk concept can remain effective, guiding one's inferences. So, conceptual engineers might not be able to really change how even a limited group of people think in a specific context, without changing the relevant folk concepts. And changing folk concepts requires converting a large population.

So far, I have argued that certain projects of conceptual engineering (those that non-consultatively try to improve patterns of inference and attention) involve paternalistic cognitive engineering. This gives us a *pro tanto* reason against these projects: the target agents could reasonably regard the projects as violating their intellectual sovereignty. Before I go on to discuss whether this concern can be outweighed by reasons in favor of engineering, let us briefly distinguish the worry raised here from other ethical concerns about conceptual engineering.

First, one might have different antipaternalist concerns. I have been discussing a broadly Kantian antipaternalist worry, one that appeals to autonomy in the sense of sovereignty. But there are also broadly Millian antipaternalist concerns, which revolve around the paternalists' fallibility. For example, when we do not consult the target agents, we may be mistaken about what is good for them, and so end up harming rather than helping them. Simpson discusses this worry in the context of epistemic paternalism in education:

> The worry that's lurking, for the epistemic paternalist, is that if you aren't being consultative, it's hard to make a person's inquiry more successful, because the aims of her inquiry, and thus what would qualify as a success in it, evolve as the inquiry develops, in a way that the paternalizer isn't well placed to keep track of. Whether you are epistemically benefiting someone depends on whether you're helping them

gain interest-relevant beliefs, knowledge, or understanding. (Simpson 2021: 98)

This consideration is also pressing for conceptual engineers trying to improve other people's patterns of attention. How can the conceptual engineers know, without consulting the target agents, that they are really directing attention to the most interest-relevant matters? And a similar worry, also related to the paternalists' epistemic situation, arises for inference-shaping conceptual engineering that is supposed to facilitate true thoughts and suppress false thoughts: the conceptual engineer might be mistaken about the truth and falsehood of the thoughts in question.

Such worries about the paternalists' epistemic situation revolve around Mill's primary concern regarding paternalism: the paternalists' fallibility. For example, Mill writes that "if any opinion is compelled to silence, that opinion may, for aught we can certainly know, be true. To deny this is to assume our own infallibility" (Mill 1999/1859: 97). In this spirit, one might insist that we should not use conceptual engineering to reduce the occurrence of false thoughts and to encourage true thoughts: for all we can certainly know, these thoughts may not be false and true, respectively. Further, conceptual engineers can be wrong about other relevant matters, like which inferences the proposed language policies would end up promoting or discouraging; and more generally, they might be unable to predict the consequences of their benevolent interventions. Marques's (2020) discussion of the ethics of conceptual engineering focuses on fallibility worries of the latter sort, worries that well-intentioned engineers can give rise to "meaning perversions", such as the shifts in the meanings of terms like "heroic" and "enemy of the people" under oppressive propagandistic regimes.

While such Millian worries about conceptual engineers' fallibility are also important and worth further discussion, they are distinct from my sovereignty-based concern. We may combine both worries, however, borrowing Feinberg's words: "By and large, a person will be better able to achieve his own good by making his own decisions, but even where the opposite

is true, others may not intervene, for autonomy is even more important a thing than personal well-being" (Feinberg 1983: 460).

Finally, there are also ethical concerns about conceptual engineering that have nothing to do with paternalism. Naturally, one might be concerned about engineering that aims at the *engineers'* good, not at the target agents' good. For example, Queloz and Bieber (2021) argue that making conceptual engineering easy to implement by somehow institutionalizing it—perhaps through a Ministry of Conceptual Engineering—would be illiberal and undemocratic, opening the door to abuses and allowing those in power to engineer consent, instead of securing genuine consent. I have left such concerns about malevolent or selfish engineering aside here, focusing on the less obvious ethical concerns that arise for benevolent conceptual engineering. Granted, as Shields (2021) points out, the current discussion on conceptual engineering might well give one the impression that shaping concepts is typically benevolent—more precisely, that it is motivated by an interest in establishing the correct or best concepts to use, rather than the shapers' selfish interests; and that impression is mistaken. Nevertheless, I focus on benevolent conceptual engineering in this paper.

4. Justifying paternalistic conceptual engineering

So far, I have identified a *pro tanto* problematic variety of epistemic paternalism—paternalistic cognitive engineering—and placed some projects of conceptual engineering within this category. The *pro tanto* reason against such engineering is that the target agents could reasonably regard it as a violation of their intellectual sovereignty. But are there reasons in favor of these projects that could outweigh the sovereignty-based reason against them? I will consider two avenues for justifying the projects: first, by appeal to the target agents' epistemic good alone, and second, by appeal to the epistemic improvement's broader social impact.

First, one might appeal to the target agents' own epistemic good with the sort of move that the defenders of nudging (e.g., Thaler and Sunstein 2008) often make: people's choices are in any case influenced by how information is presented or how the environment is arranged. For example, people will in any case tend to choose the food at eye level. So, if we refrain from interfering paternalistically, their choice will instead be influenced by chance or possibly by actors who do not have their interests in mind. The same point can be made about conceptual engineering. Concepts shape patterns of inference and attention, with or without our intervention; and unless we interfere for the concept users' own epistemic good, they will be influenced by how their concepts are shaped by chance or possibly by malevolent actors. One might say, then, that if we leave people to govern their own concepts, we in effect leave them at the mercy of other (intentional and unintentional) forces that are not benevolent like us. So, even if there is a sovereignty-based reason against engineering, there is also a reason to engineer, arising from concern for the target agents' epistemic well-being.

However, I doubt that this consideration alone can outweigh the sovereignty-based reason against non-consultative conceptual engineering. The engineers do not face a choice between engineering non-consultatively and doing nothing. They also have the option of promoting awareness of how concepts influence belief-formation—how concepts in general and specific concepts make certain things salient and put our thoughts on certain tracks—and offering their own suggestions for conceptual revision, along with the reasons. The engineers could thus empower people to exercise control over their mental lives, instead of just taking the matter into their own hands. It seems, then, that a compelling case for non-consultative engineering would need weightier reasons.

Another possible defense of non-consultative conceptual engineering (again, by appeal to the target agents' own epistemic interests, broadly construed) is the following. While some people may wish not to have their concepts non-consultatively managed by the engineers,

others might not mind and might even appreciate it; and that preference is also reasonable. One might well prefer not having to consider the justifications that consultative engineers offer for their proposals. This takes time and effort that one could spend on other matters, while benevolent engineers non-consultatively look out for one's interests in shaping the community's shared conceptual repertoire. For people who would rather just go with the flow when it comes to concepts, benevolent non-consultative conceptual engineers conveniently make sure that the flow takes them in a good direction.

It is unclear, again, whether this consideration outweighs the sovereignty-based concern. Even if many or most people take such a lax attitude toward having their intellectual lives controlled for their own epistemic good, the engineers cannot legitimately assume that all their target agents have this attitude. For an analogy: perhaps many or even most people (reasonably) appreciate non-consultative hugging and feel annoyed about being consulted, while some (also reasonably) regard non-consultative hugging as a violation of their bodily sovereignty. Can we then just go ahead and non-consultatively hug someone whose preference we do not know? It seems that there is a greater presumption against violating bodily sovereignty than against disrespecting an unspoken wish to be non-consultatively hugged. (The latter is plausibly not a sovereignty violation, if sovereignty means primarily a right *not* to be treated in certain ways without our consent, in the relevant domain, rather than a right *to* be treated in our preferred ways, in that domain.) So, perhaps we should err on the side of caution by refraining from the hug, or by consulting first; but I leave this (as well as the potentially relevant differences between non-consultative hugging and non-consultative conceptual engineering) for future discussion.

So far, I have discussed how the target agents' own epistemic good could outweigh the sovereignty-based concern. One might also appeal to the broader social impact of the epistemic improvement. For example, Ahlstrom-Vij (2018) justifies withholding evidence from jurors

like this: "[A]ny legitimate moral claim on the part of the jury not to be interfered with is trumped by considerations about the welfare of others, such as the defendant and those allegedly wronged by the defendant" (271). Indeed, the ultimate motivation behind epistemic paternalism often has to do with the interests of people other than the target agents, or with the non-epistemic flourishing of society more generally (including the target agents). Likewise, the motivation behind conceptual engineering is often that our concepts have far-reaching societal consequences beyond the immediate cognitive effects for the concept users.[4]

Pursuing this line of defense, one might further point out that we can almost always expect our epistemic success and failure to have broader social impact in some roundabout way. For example, Clifford (1999/1877) argued that false beliefs tend to spread by various mechanisms, and this can have far-reaching consequences for society. An apparently insignificant false belief can lead one to revise their other beliefs and to interpret incoming evidence based on the false belief, resulting in deterioration of the quality of a bigger portion of their beliefs. And this in turn can result in harmful behavior based on the false beliefs or the spreading of those beliefs to others via testimony. So, when someone's belief-formation goes astray, we can typically expect adverse effects for others, regardless of the content of the belief. The lesson that Clifford takes from this is that we bear a general responsibility to other people to govern our beliefs well. But one might also draw the further conclusion that we can non-consultatively interfere to secure the quality of our fellow humans' belief-formation, even in the absence of specific reason to suspect that the beliefs in question are consequential for us.

However, it is dubious that these Cliffordian considerations give us a free pass to do just anything to improve other people's epistemic state. Consider Bullock's (2018: 442) example of secretly playing lectures on quantum mechanics to someone while they sleep. Even if this

[4] For example, Burgess and Plunkett motivate evaluating and improving concepts with the idea that our concepts' "non-conceptual consequences are pervasive and profound" (Burgess and Plunkett 2013, 1096–1097).

would be an effective way to teach quantum mechanics, we cannot justify this intervention merely on the Cliffordian grounds that the person's ignorance of quantum mechanics could well end up harming others in some roundabout way that we cannot yet predict; or that their knowledge is likely to benefit society in some roundabout way and the sovereignty-violation is thus justified.

So, on one hand, it is plausible that the broader social impact of the epistemic improvement can speak in favor of paternalistic cognitive engineering (including conceptual engineering) and outweigh the sovereignty-based reason against these practices. On the other hand, as the quantum mechanics example shows, the justification cannot be merely that epistemic improvement typically has such broader social impact. There is no escape from looking at the specifics of a given paternalistic project. To make this point clearer, let us look at two possible projects of conceptual engineering: one that is plausibly justified and another where this is less plausible.

(1) *Discouraging rape metaphors.* Suppose that some of Fraser's readers become convinced that the relevant rape metaphors contribute to mistaken ways of thinking about rape—for example, that rape victims should have put up more of a fight and that there is no great injustice in rape. These readers join forces against such metaphors. They use arguments, for example, publishing newspaper articles and blog posts to popularize and build on Fraser's case. Recognizing that this alone is not enough, they also unfriend people who use such metaphors on social media and sanction them in other informal ways, contributing to the emergence of a social norm against the metaphors. Further, they try to get social media influencers on board and do their best to convince newspaper and journal editors to remove the rape metaphors.[5] The

[5] Nimtz (2021) and Thomasson (2021) both suggest that we can engineer social norms to engineer concepts, and that wielding trendsetters' (authoritative speakers') influence has an important role in engineering social norms.

conceptual activists know that many of the targeted concept users would not be convinced by their arguments and that the majority will not even be aware of the arguments.

A *pro tanto* reason against this project is that the target agents could reasonably regard it as a violation of their intellectual sovereignty. However, there are also powerful reasons to use non-consultative cognitive engineering here and not just directly address the mistaken ways of thinking about rape, or engineer concepts entirely consultatively. It is very difficult to change ingrained misogynous attitudes by arguments; and for related reasons, it is very difficult to rationally convert many of those prone to using rape metaphors. Further, these mistaken ways of thinking cause further suffering for victims and could contribute to failures to deter rape—so, the epistemic improvement here demonstrably and not just potentially has important social impact. Moreover, it is beyond reasonable doubt that the discouraged ways of thinking about rape are indeed mistaken and harmful; and it is also very plausible that the rape metaphors indeed perpetuate these mistaken and harmful ways of thinking. And there are even further reasons against using rape metaphors: these metaphors are often triggering and insulting for victims. It is thus plausible that the activism against rape metaphors is ultimately justified, despite the sovereignty-based reason against it.

(2) Revising gender terms. A group of people becomes convinced that revising gender terms as suggested by Haslanger (2000) would result in a gestalt shift that makes gender inequality more salient and thereby motivates progressive action. They want as many people as possible to adopt the revised concepts. The activists use the methods described in the previous example, that is, rational persuasion combined with shaping social norms on the use of gender terms. They also try to convince dictionary compilers to mention the proposed definitions; and they lobby for various institutions to recommend or require this usage in their language guidelines. When people object to

> the proposals (for example, as unnecessarily confusing), they dismiss it as conservative backlash. They adopt a policy of speaking as if the terms already had the revised meanings and celebrate the resulting confusion as facilitating critical reflection on the current meanings of the terms.

The engineers, again, might insist that the worthy goal of gender equality outweighs the sovereignty-based reason against the project. However, various features of the project make that response less convincing here. Granted, gender equality is a worthy aim, and it is somewhat plausible that the conceptual revision would indeed contribute to that aim, by guiding attention to gender inequality. However, these engineers' methods are less focused on rationally persuading the target agents; and they go further in trying to wield institutional power and social sanctions. One might also question whether the heightened attention to gender inequality, across all contexts where gender terms come up, best serves everyone's epistemic interests. There might be other valuable ways of thinking about the relevant aspects of the world—one might want to emphasize gender identity rather than gender oppression, for example (see Jenkins 2016). In sum, these Haslanger-inspired engineers should seriously consider other ways of promoting gender equality or at least pursue the conceptual revision more consultatively. (I do not wish to suggest, of course, that Haslanger herself envisioned the project as described here.)

There is no single thing that decisively sets this project apart from the previous one; various features matter. For example, it is relevant that the activists who try to eliminate rape metaphors are protecting people from clearly harmful false beliefs, whereas those trying to revise gender terms are guiding attention to some important concerns among others; accordingly, there is more room for reasonable disagreement in the latter case. The additional reasons for discouraging rape metaphors are also relevant and contribute to the justification of that project; and so on. Perhaps some would nevertheless consider the project of revising gender terms to

be justified as well, as long as it significantly contributes to gender equality. I do not insist that my judgment on the cases is necessarily correct. The main aim here is just to show that various details matter for assessing whether an engineering project is ultimately justified, despite the sovereignty-based reason against it. It is not enough, then, to make the Cliffordian point that an individual's epistemic well-being typically has broader social impact; and it is also not enough to just gesture toward some worthy aim (such as gender equality) that supposedly outweighs infringing on intellectual sovereignty. The project's specific aims, means, and social context matter.

5. Concluding remarks

I argued that non-consultatively engineering concepts to improve inferential and attentional dispositions falls into a *pro tanto* problematic category of epistemic paternalism: paternalistic cognitive engineering. This category arguably also includes certain kinds of nudging and evidence suppression. There is a reason not to engage in paternalistic cognitive engineering: the target agents could reasonably take it to violate their intellectual sovereignty. They could prefer to be persuaded, informed, and educated (which may also include consultative conceptual engineering) rather than non-consultatively engineered, since this would give them more control over their belief-formation. I assumed without much argument that it is reasonable to be rather protective of one's intellectual domain and to oppose exercises of control over that domain when methods like explicit persuasion are in principle available. However, I acknowledge that a laxer attitude toward such benevolent exercises of control is *also* reasonable.

The sovereignty-based concern raised here is only a *pro tanto* reason against such conceptual engineering projects. I discussed two ways of justifying these projects despite the sovereignty-based reason against them. First, one could insist that concern for the target agents' own epistemic well-being outweighs the sovereignty-based worry. If we refrain from

engineering, people's inferential and attentional patterns will be shaped by chance and possibly by malevolent actors. Further, some target agents might even appreciate the non-consultative benevolent engineering, far from regarding it as sovereignty-violating. I left this avenue for justification open but expressed some reservations. Second, one could propose that the broader social impact of the epistemic improvement outweighs the sovereignty-based concern. I argued that this needs to be assessed case by case, since there are various relevant differences between projects of conceptual engineering.

Acknowledgments. I am thankful to the participants of the CONCEPT group's brown bag seminar at the University of Cologne, where I presented a previous version of the paper—especially Sven Bernecker, Sofia Bokros, Adam Bricker, Thomas Grundmann (who also provided helpful written comments on the paper), Luis Rosa, César Schirmer dos Santos, and Paul Silva. The constructive comments of the referees for the *Journal of the American Philosophical Association* and other journals have also improved the paper significantly. The research was generously supported by the Alexander von Humboldt Foundation.

References

Ahlstrom-Vij, Kristoffer (2013) *Epistemic Paternalism: A Defense.* Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Ahlstrom-Vij, Kristoffer (2018) 'Epistemic Paternalism'. In Kalle Grill and Jason Hanna (eds.), *The Routledge Handbook of the Philosophy of Paternalism* (London and New York: Routledge), pp. 261–273.

Brandom, Robert B. (1994) *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Cambridge, MA: Harvard University Press.

Bullock, Emma (2018) ‘Knowing and Not-Knowing for Your Own Good: The Limits of Epistemic Paternalism’. *Journal of Applied Philosophy*, 35 (2), 433–447.

Burgess, Alexis, and David Plunkett (2013) ‘Conceptual ethics I’. *Philosophy Compass*, 8 (12), 1096–1097.

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

Castro, Clinton, Adam Pham, and Alan Rubel (2020) ‘Epistemic Paternalism Online’. In Amiel Bernal and Guy Axtell (eds.), *Epistemic Paternalism: Conceptions, Justifications and Implications* (London: Rowman & Littlefield), pp. 29–43.

Chalmers, David (2020) ‘What is conceptual engineering and what should it be?’. *Inquiry*, DOI: 10.1080/0020174X.2020.1817141.

Clark, Andy and David Chalmers (1998) ‘The Extended Mind’. *Analysis*, 58 (1), 7–19.

Clifford, William Kingdon (1999/1877) *The Ethics of Belief and Other Essays.* Amherst, NY: Prometheus Books.

Cohen, Stewart (2016) ‘Theorizing About the Epistemic’. *Inquiry*, 59 (7–8), 839–857.

Deutsch, Max (2020a) ‘Speaker’s Reference, Stipulation, and a Dilemma for Conceptual Engineers’. *Philosophical Studies*, 177, 3935–3957.

Deutsch, Max (2020b) ‘Trivializing conceptual engineering’. *Inquiry*, DOI: 10.1080/0020174X.2020.1853343.

Feinberg, Joel (1983) ‘Autonomy, Sovereignty, and Privacy: Moral Ideals in the Constitution’. *Notre Dame Law Review*, 58 (3), 445–492.

Fischer, Eugen (2020) ‘Conceptual control: on the feasibility of conceptual engineering’. *Inquiry*, DOI: 10.1080/0020174X.2020.1773309.

Fraser, Rachel Elizabeth (2018) 'The Ethics of Metaphor'. *Ethics*, 128, 728–755.

Goldman, Alvin I. (1991) 'Epistemic Paternalism: Communication Control in Law and Society'. *The Journal of Philosophy*, 88 (3), 113–131.

Grundmann, Thomas (2021a) 'The Possibility of Epistemic Nudging'. *Social Epistemology*. DOI: 10.1080/02691728.2021.1945160.

Grundmann, Thomas (2021b) 'The Possibility of Epistemic Nudging: Reply to My Critics'. *Social Epistemology Review and Reply Collective*, 10 (12), 28-35.

Haslanger, Sally (2000) 'Gender and Race: (What) Are They? (What) Do We Want Them To Be?'. *Noûs*, 34 (1), 31–55.

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

Hausman, Daniel M. and Brynn Welch (2010) 'Debate: To Nudge or Not to Nudge'. *The Journal of Political Philosophy* 18 (1), 123–136.

Hazlett, Allan (2016) 'What Does 'Epistemic' Mean?'. *Episteme*, 13 (4), 539–547.

Isaac, Manuel Gustavo (2020) 'How to Conceptually Engineer Conceptual Engineering?'. *Inquiry*, DOI: 10.1080/0020174X.2020.1719881.

Isaac, Manuel Gustavo (2021a) 'Which Concept of Concept for Conceptual Engineering?'. *Erkenntnis*, https://doi.org/10.1007/s10670-021-00447-0

Isaac, Manuel Gustavo (2021b) 'Broad-spectrum conceptual engineering'. *Ratio*, 34 (4), 286–302.

Jackson, Elizabeth (2021) 'What's Epistemic About Epistemic Paternalism?'. In Jonathan Matheson and Kirk Lougheed (eds.), *Epistemic Autonomy* (New York: Routledge), pp. 132–149.

Jenkins, Katharine (2016) ‘Amelioration and Inclusion: Gender Identity and the Concept of Woman’. *Ethics*, 126, 394–421.

Koch, Steffen (2021) ‘Engineering what? On concepts in conceptual engineering’. *Synthese*, 199, 1955–1975.

Levy, Neil (2019) ‘Nudge, Nudge, Wink, Wink: Nudging is Giving Reasons’. *Ergo*, 6 (10), 281–302.

Levy, Neil (2021a) ‘Nudging is Giving Testimony: A Response to Grundmann’. *Social Epistemology Review and Reply Collective*, 10 (8), 43–47.

Levy, Neil (2021b) ‘Predictably Rational: A Further Response to Grundmann’. *Social Epistemology Review and Reply Collective*, 10 (12), 75–79.

Levy, Neil (2022) *Bad Beliefs: Why They Happen to Good People*. Oxford: Oxford University Press.

Machery, Edouard (2017) *Philosophy Within Its Proper Bounds.* Oxford: Oxford University Press.

Machery, Edouard (2021) ‘A new challenge to conceptual engineering’. *Inquiry*, DOI: 10.1080/0020174X.2021.1967190.

Marques, Teresa (2020) ‘Amelioration vs Perversion’. In Teresa Marques and Asa Wikforss (eds.), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variability* (Oxford: Oxford University Press), pp. 260–284.

McKenna, Robin (2020) ‘Persuasion and Epistemic Paternalism’. In Amiel Bernal and Guy Axtell (eds.), *Epistemic Paternalism: Conceptions, Justifications and Implications* (London: Rowman & Littlefield), pp. 91–106.

Medvecky, Fabien (2020) ‘Epistemic Paternalism, Science, and Communication’. In Amiel Bernal & Guy Axtell (eds.), *Epistemic Paternalism: Conceptions, Justifications and Implications* (London: Rowman & Littlefield), pp. 79–89.

Mill, John Stuart (1999/1859) *On Liberty*. Edited by Edward Alexander. Peterborough: Broadview Press.

Nimtz, Christian (2021) ‘Engineering concepts by engineering social norms: solving the implementation challenge’. *Inquiry*, DOI: 10.1080/0020174X.2021.1956368.

Pinder, Mark (2021) ‘Conceptual Engineering, Metasemantic Externalism and Speaker-Meaning’. *Mind*, 130 (517), 141–163.

Queloz, Matthieu, and Friedemann Bieber (2021) ‘Conceptual Engineering and the Politics of Implementation’. *Pacific Philosophical Quarterly*, https://doi.org/10.1111/papq.12394.

Riley, Evan (2017) ‘The Beneficent Nudge Program and Epistemic Injustice’. *Ethical Theory and Moral Practice*, 20, 597–616.

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

Schwitzgebel, Eric (2021) ‘The Pragmatic Metaphysics of Belief’. In Cristina Borgoni, Dirk Kindermann, and Andrea Onofri (eds.). *The Fragmented Mind* (Oxford: Oxford University Press), pp. 350–375.

Shields, Matthew (2021) ‘Conceptual Domination’. *Synthese* 199, 15043–15067.

Simpson, Robert Mark (2021) ‘Norms of Inquiry, Student-Led Learning, and Epistemic Paternalism’. In Jonathan Matheson and Kirk Lougheed (eds.), *Epistemic Autonomy* (New York: Routledge), pp. 95–112.

Sterken, Rachel (2020) ‘Linguistic Intervention and Transformative Communicative Disruptions’. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds).

Conceptual Engineering and Conceptual Ethics (Oxford: Oxford University Press), pp. 417–434.

Thaler, Richard H. and Cass R. Sunstein (2008) *Nudge: Improving Decisions about Wealth, Health, and Happiness.* New Haven: Yale University Press.

Thomasson, Amie (2021) ‘Conceptual engineering: when do we need it? How can we do it?’. *Inquiry*, DOI: 10.1080/0020174X.2021.2000118.