Philosophy & Technology (2022) 35:71
https://doi.org/10.1007/s13347-022-00565-9

RESEARCH ARTICLE

Technology as Driver for Morally Motivated Conceptual Engineering

Herman Veluwenkamp[1] · **Marianna Capasso**[2] · **Jonne Maas**[1] · **Lavinia Marin**[1]

Received: 11 January 2022 / Accepted: 12 July 2022

Abstract

New technologies are the source of uncertainties about the applicability of moral and morally connotated concepts. These uncertainties sometimes call for conceptual engineering, but it is not often recognized when this is the case. We take this to be a missed opportunity, as a recognition that different researchers are working on the same kind of project can help solve methodological questions that one is likely to encounter. In this paper, we present three case studies where philosophers of technology implicitly engage in conceptual engineering (without naming it as such). We subsequently reflect on the case studies to find out how these illustrate conceptual engineering as an appropriate method to deal with pressing concerns in the philosophy of technology. We have two main goals. We first want to contribute to the literature on conceptual engineering by presenting concrete examples of conceptual engineering in the philosophy of technology. This is especially relevant, because the technologies that are designed based on the conceptual work done by philosophers of technology potentially have crucial moral and social implications. Secondly, we want to make explicit what choices are made when doing this conceptual work. Making explicit that some of the implicit assumptions are, in fact, debated in the literature allows for reflection on these questions. Ultimately, our hope is that conscious reflection leads to an improvement of the conceptual work done.

Keywords Technology · Conceptual Engineering · Moral Conflict · Freedom · Critical Thinking · Control

✉ Herman Veluwenkamp
h.m.veluwenkamp@tudelft.nl

[1] Delft University of Technology, Delft, Netherlands

[2] Sant'Anna School of Advanced Studies, Pisa, Italy

Published online: 28 July 2022

1 Introduction

New technologies are the source of puzzlement and considerable moral uncertainty. How should we think about the technological issues, which principles apply, which values are salient, and what do we owe to those affected by the negative consequences of our innovations? In addition to the problems about deciding which ethical principles and theories apply, there is a problem of how to conceptualise the parts of the world that we are dealing with. The introduction of the mechanical ventilator and the early pregnancy test have, for example, caused strong disagreement about the application conditions of what philosophers term 'thick descriptive concepts' such as *death* and *pregnancy* (Baker and of P. R. 2019; Leavitt, 2006). These technology-induced uncertainties about the applicability of moral and morally connotated concepts are becoming very prominent. The worldwide adoption of CRISPR-CAS genetic engineering techniques has given rise to fierce moral and legal debates in Europe about what counts as a 'natural way' of altering the genome.[1] Can one have a thousand friends online, or do we conclude that those who claim such an impressive circle of friends must have a radically different conception of friendship? Blockchain and quantum encryption can mathematically guarantee that parties will comply with norms, promises and contracts. Now we can make our interaction 'failsafe'; have we therefore established *trust* between parties? Many of the problems we have with technology are of this type and they call for conceptual engineering. In this paper, we analyse examples of how modern technology calls for conceptual engineering. Conceptual engineering, however, is a field that has only recently received considerable attention, and many important methodological questions remain open (Eklund 2021). We use our analysis of technology-induced examples of conceptual engineering to see how these methodological questions are answered in the technological domain.

Conceptual engineering as we understand it is the design, implementation and evaluation of concepts (Chalmers, 2020, p. 2).[2] This practice would not be justified if we had already employed the best possible concepts,[3] but there are multiple reasons for doubting that this is so. The analytical tradition in philosophy is based on the idea that we should not just make do with the concepts we have been handed. It is shown — sometimes in great detail — how our language leads us astray, into paradoxes, puzzles, contradictions, absurd conclusions, or plain nonsense. Even if our preferred frameworks of philosophical concepts are generally fit for purpose (as explanatory tools), we still might find concepts or even subsets of concepts that are defective in some ways. A concept can be defective in different ways. Herman Cappelen (2018) identifies two major variants of conceptual deficiency. Firstly, the semantic value of a concept can be defective. Early emotivists, for example, held that our normative concepts are meaningless and should therefore be revised or even abandoned. But the inconsistency of

[1] We would like to thank Jeroen van den Hoven for suggesting many of the examples mentioned in this section.

[2] As is usual with a term of art as 'conceptual engineering', not everyone uses this term in the same way. Cappelen (2018, p. 3), for example, agrees that conceptual engineering should be engaged with design, implementation and evaluation, but takes the proper target of the practice to be representational devices.

[3] The 'best possible concept' should be understood as the all-things-considered best possible concept. See for a related discussion Eklund (2012).

a concept is also used as a reason for revising the concept. *Truth*, *freedom*, *knowledge* and *race* have all been called inconsistent or incoherent, and alternative conceptions have been proposed. The second kind of conceptual deficiency Cappelen identifies is that the concept is morally, politically or socially problematic. The concept *marriage*, for example, is not semantically defective. However, if it excludes same-sex couples, then that might have objectionable moral, political and social consequences. One of the consequences that motivated a conceptual change, in this case, is that people could come to see same-sex relationships as inferior. Sally Haslanger's ameliorative projects are also explicitly politically motivated. Her stated goal is the elimination of what she calls women. She believes that 'it is part of the project of feminism to bring about a day when there are no more women (though, of course, we should not aim to do away with females!)' (Haslanger, 2000, p. 46).

The kind of deficiency we are interested in is of the second kind. In the case of semi-autonomous cars, for example, we see that the application of our ordinary conceptions of control and responsibility causes moral uncertainty. It is not clear whether they introduce responsibility gaps (Matthias, 2004; Sparrow, 2007) and how this affects public policy options regarding transportation. Moreover, even if they do not introduce responsibility gaps, it is not clear what the proper distribution of responsibility should be, as Google and Uber's different reactions to a semi-autonomous car crash show (Nyholm, 2018). In this context, it is not the case that there is moral uncertainty because we do not know which moral theory to apply — as Nickel et al. (2020) had pointed out. It is also not the case that the application conditions of a thick descriptive concept have changed. Instead, we see a conflict between the technology, our moral judgements and the relevant moral concepts. One way of resolving this conflict is by proposing new or revised moral concepts or conceptions.

We call *morally motivated conceptual engineering* the process of looking for the best conceptions triggered by a conflict between moral judgements and new contexts. As this kind of conceptual engineering is not triggered by a semantic defect in the conceptions used, it is of the second, ameliorative kind. We see that this kind of morally motivated conceptual engineering steadily increases in importance. That this work is a kind of conceptual engineering is, however, not often recognized. We take this to be a missed opportunity, as a recognition that different people are working on the same kind of project can help solve methodological questions that one is likely to encounter. In this paper we have two main goals. We first want to contribute to the literature on conceptual engineering by presenting concrete examples of conceptual engineering in the philosophy of technology. This is especially relevant, because the technologies that are designed based on the conceptual work done by philosophers of technology[4] are morally and socially significant (see, e.g. Friedman & Hendry, 2019 and van den Hoven, 2013; 2017). Secondly, we want to make explicit what methodological choices are made when doing this conceptual work. Making explicit what methodology we are employing allows for a conscious

[4] We use the terms 'philosophers of technology' and 'philosophy of technology' in a broad sense. It applies respectively to scholars that critically reflect on the use of technology in society and the field in which these scholars work, though this need not be limited to 'philosophers' alone.

reflection on this methodology. Ultimately, our hope is that conscious reflection leads to an improvement of the used methods.

To accomplish this all, we proceed as follows. We first provide a framework for understanding conceptual work done by philosophers of technology as conceptual engineering (Sect. 2). We then present three case studies through the lens of this framework (Sect. 3). Finally, we reflect on the case studies to find out how these philosophers deal (or should deal) with some of the pressing questions in the conceptual engineering literature. Questions we discuss include the question of what method to use for doing conceptual engineering and how to determine when a conceptual change changes the original topic (Sect. 4).

2 Conceptual Engineering

Let us first introduce some terminology. We distinguish between a concept and its conception(s). Concepts stand to conceptions in a one to many relation, that is, different conceptions can be *of* a concept. In his *A Theory of Justice* (1999), John Rawls introduces this distinction as follows:

> Men disagree about which principles should define the basic terms of their association. Yet we may still say, despite this disagreement, that they each have a conception of justice. That is, they understand the need for, and they are prepared to affirm, a characteristic set of principles for assigning basic rights and duties and for determining what they take to be the proper distribution of the benefits and burdens of social cooperation. Thus it seems natural to think of the concept of justice as distinct from the various conceptions of justice and as being specified by the role which these different sets of principles, these different conceptions, have in common (Rawls 1999, p. 5)

Rawls seems to have in mind here that there is a specific role or function (providing 'principles for assigning rights and duties') and that different conceptions perform this function to a certain extent. The relevant question now becomes which conception performs this function best, given the context under consideration. And, of course, Rawls famously argued that in the context of western liberal democracies, justice as fairness is the best conception of justice.

In line with this take on the concept/conception distinction, we propose to understand concepts and conceptions as entities that have meanings as their content.[5] The meanings of concepts are in some sense indeterminate, and there are different

[5] Philosophers of language typically distinguish between representationalist and inferentialist theories of meanings. Representationalists define meaning in terms of what a conception purports to represent. So if two conceptions purport to represent the same things, then they are identical (e.g. an equilateral triangle and an equiangular triangle). Inferentialists, on the other hand, individuate conceptions in terms of conceptual or inferential role (which can be spelled out in terms of what one is disposed of or in terms of what one is normatively committed to infer). The fundamental distinction between those theories is one of explanatory priority, as most representationalists hold that conceptions have an inferential role and inferentialists maintain that conceptions purport to represent things. We are most attracted to inferentialist theories of meaning, but for this paper, we opt to remain neutral on this issue.

ways of making a concept precise. If we make a concept precise without a change of topic, then we have a conception of that concept. *Justice as Fairness* can now be understood as one way of making *Justice* precise. Not all ways of making a concept more precise will be without a change of topic. If we propose, for example, a conception which identifies justice with being a blue sky, then we will no longer be talking about our old concept of justice anymore. Two conceptions are of the same concept if they are similar enough. In some cases, two conceptions are similar enough if they play the same role, as the above interpretation of Rawls suggests. However, to be maximally inclusive in our definition of the relation between concepts and conceptions, we leave open the possibility that other factors determine what counts as similar enough (see, e.g. Lalumera, 2014 for different interpretations of the concept/conception relation).

We understand the practice of conceptual engineering as trying to find out what the correct conception of a concept is in a specific context. Although conceptual engineering has always been an important part of the analytic philosophical tradition, more attention has recently been paid to a systematic approach to this method. Indeed, as such, the field is heterogeneous, and philosophers differ on several questions that greatly influence how we might comprehend the core tenets of conceptual engineering. There is disagreement about what exactly it is that is being engineered when we engage in conceptual engineering: e.g. concepts (Haslanger, 2000), representational devices (Cappelen, 2018; Haslanger, 2006), intentions and extensions (Deutsch, 2020), inferential roles (Eklund 2021) or commitment and entitlement structures (Löhr, 2021). It is not our goal to settle this debate in this paper, and we will assume in this paper that conceptual engineering concerns concepts and conceptions.

It is important to note that when we say 'correct', we do not mean to ask which conception we are currently using, that is, we are not primarily engaging in conceptual descriptive analysis that tracks and explicates our ordinary understanding of concepts. Instead, we are asking the question which conception is best in line with an ameliorative inquiry, i.e. which conception would grasp best our legitimate and critical purpose of evaluating the moral, political and social dimensions of current specific tech-mediated contexts. This distinction is important, because often conceptual change is called for, exactly because there is something defective with the conceptions we are currently using, since those latter might fail to track all the relevant moral or political dimensions that a critical analysis on social reality should take into account.

Given this way of understanding conceptual engineering in the philosophy of technology, there are several open questions. First there is a methodological question: What is it that makes a conception best? In the literature about conceptual engineering, there are two main candidates: a metaphysical and a pragmatic approach (Thomasson, 2020). It is, however, not evident which method is most suited for philosophers of technology. Therefore, we do not want to commit ourselves to one particular methodology from the outset. Our goal is rather suggesting an analysis on how a comprehensive methodology or a model can be advanced for changing or adapting concepts in response to current political, moral and societal implications of emerging technologies, and through which we can also evaluate how technologies

themselves, especially emerging ones such as AI or other autonomous and intelligent systems, challenge the meaning of our concepts. Therefore, this analysis is a means to bridge the gap between two lines of inquiry: conceptual engineering methodologies (Haslanger 2020; Cappelen, 2018) and philosophy of technology studies that deal with the political, social, moral implications of technological innovation and the related transformative force (Van de Poel, 2020).[6] We do this by analysing the three case studies and by establishing which of the methodologies described in the literature corresponds best with the work done in the field. A second, prominent question when engaging in conceptual engineering is whether we are changing the topic when a new conception is proposed. This is an accusation that is sometimes made, and it is unclear what the best response to such a challenge is. The third question relates to two different views about conceptions that are sometimes defended: conception relativism and conception imperialism. The relativists hold that no conception is better than another, while the imperialist holds that if a conception is the best conception of a concept in one context, it is also best in other contexts. What we are interested in is if these views are reflected in the conceptual work that is done by philosophers of technology. The final question we will assess is whether we always have to design new concepts or whether there are other options.

3 Cases

In this section, we explore three case studies. We selected these case studies since these are the cases that in our critical examination of tech-mediated scenarios present a gap between operative concepts and our practical purposes of better framing moral, social, and political implications of emerging technologies. Moreover, they are the object of current debates in philosophy of technology, but still no consensus has been reached yet on their conceptual definition or on their normative status. When we describe these cases, we will use the term 'operative conception' to refer to the conception that is under investigation. This is not necessarily the dominant conception, but it is the conception that is used by people discussing these issues and which leads to the conflict. We will use the term 'target conception' to denote the conception that we think is better in the relevant context.

These case studies are analysed in this way: first, we discuss operative conceptions of control, critical thinking and freedom; second, we critically examine emerging technologies' moral, social, political implications; then, at this point, we indicate a misalignment or mismatch, since the operative conceptions neglect or undertheorize properties and implications individuated in our critical analysis. Thus, given the moral intuitions we have on tech-mediated scenarios that we critically examine, we finally introduce a target conception, i.e. a conception we should be using, given our purposes and goals in that inquiry, and that comes apart from the operative or ordinary ones and fits better the candidate meaning in tech-mediated scenarios.

It is helpful to make explicit what methodology we employ when we analyse these cases. The cases are all descriptions of debates that are started (or intensified) because of technological innovations. We analyse these cases to convince the reader

[6] In Sect. 4.1, we provide a more detailed overview of the different methodologies.

that the best way to understand these debates is to view them as instances of conceptual engineering. We aim to achieve this by showing how the framework from the previous section can be applied to the central questions in the debate and how this helps understanding which moves are and which moves are not legitimate.

This does not mean that all disputants agree with us that they are engaging in conceptual engineering. On the contrary, if that were the case, there would be no need to write this article. Instead, we want to convince people that it is useful to see these debates as morally motivated conceptual engineering. We therefore first discuss the central concept and an operative conception that is (implicitly) used in the arguments surrounding the case. We then show how this conception leads to the moral judgements that are taken to be problematic. Finally, given such an assessment of the misalignment, we introduce a target conception, i.e. a conception we should be using, given our purposes and goals in that inquiry, one that comes apart from the operative or ordinary ones and fits better the candidate meaning in tech-mediated scenarios. This analysis can be considered as a kind of ameliorative conceptual engineering in the philosophy of technology studies/literature.[7]

3.1 Control

3.1.1 Operative Conception

It is philosophical orthodoxy to define moral responsibility in terms of an epistemic and a control condition (Rudy-Hiller, 2018), that is, a moral agent is typically taken to be responsible for an action if and only if that agent is adequately aware of the consequences of performing the action and possesses a sufficient degree of control over the action. The conception of control has been, and still is, debated with regard to its connection to determinism: 'Is control possible if the world is governed by deterministic rules?'. Apart from this fundamental philosophical question, it has been thought that the conception of control is relatively unproblematic.[8] The engineering conception of control says that as long as one is physically capable of intervening in a system's operation, one has operational control over that system.[9]

3.1.2 Conflict with Moral Judgements

Recent research in ethics of technology has insisted, however, that traditional common-sense and engineering conceptions of control should be problematized in the context of intelligent autonomous systems (Santoni de Sio and Hoven 2018; Himmelreich, 2019). One of the reasons for this is that the traditional engineering conception of control leads to verdicts that conflict with our moral judgements about

[7] Of course, it is not always legitimate to introduce a new conception. In Sect. 4, we will discuss the conditions more extensively.

[8] But see Fischer and Ravizza (1998) and Himmelreich (2019) for a philosophical discussion of control.

[9] This is also the conception of control that is endorsed by Matthias (2004) and Sparrow (2007). But note that also in mining and building destruction, we talk about 'controlled explosions'. We think we have control by making sure that an end result is achieved, but once the process is underway, we cannot possibly intervene.

semi-autonomous vehicles, for example, suppose we consider someone behind in the driver seat of a car which has adaptive cruise control switched on. As long as this person is physically able to intervene in the operation of the car, this person has operational control and is therefore potentially responsible when something goes wrong. It also leads to the claim that fully autonomous vehicles should be forbidden. The motivation for this assertion is that according to the operative conception of control there is no moral agent for whom the control condition is satisfied. Such a scenario has been called a responsibility gap: a situation where no one can be blamed for the harms of an autonomous system (Matthias, 2004).

Digital platforms are another example in the context of intelligent autonomous systems in which we can see that the operative conception of control conflicts with our moral judgements. In the documentary 'The Social Dilemma' (Orlowski, 2020), we see on full display the harm that social media companies such as Facebook and Twitter in fact do to their users. The makers of this documentary, however, seem reluctant to blame the creators of the algorithms that these companies use. Instead, these developers are the protagonists of the documentary, explicitly portrayed as ordinary people caught in a bigger game, who make nervous small talk before the actual filming starts. *They* are not in control of the algorithms! So, given the popular idea that responsibility requires control, one might think that the developers are not to blame for the harm done. However, it is also implausible to hold the users, or the algorithms themselves, responsible. Thus, one might think that a responsibility gap arises here as well. This may fuel the feeling of a technological and economic determinism, in which impersonal technological and economic forces govern society and ultimately our lives, so we may just sit on our sofa and, so to speak, enjoy the show as much as we can.

3.1.3 Target Conception

A first response to this conflict is the observation that decades of philosophy and sociology of technology have shown that technology is not determined by mysterious inscrutable forces, although there may be path dependencies. Despite their complexity, machine-learning systems often serve the interests of the companies that developed them quite well. So companies are to blame? Perhaps, but this is not the point. Instead, the goal should be to achieve a more acceptable distribution of control and responsibility in our current socio-technical landscape. Responsibility gaps may well arise in complex socio-technical systems like the social media companies, and they are a problem insofar as we know that if all stakeholders feel they have an excuse if something goes wrong, nobody has an incentive to avoid these situations. But, again, responsibility gaps are not a destiny. If they are clearly on the horizon, accompanied with a technological project, then there is a forward-looking responsibility to prevent their emergence or mitigate their effects. There are various societal stakeholders who could be addressed with such responsibility, such as designers, software developers, engineers, and regulators.

> 'Meaningful human control' (a term invented in the political debate on autonomous weapon systems) on the contrary requires more than the traditional conception of control. It requires that our interaction with the technology is designed to give us a fair capacity and opportunity to have our deepest values and interests reflected in

the behaviour of the technology. And that nobody is in the position to control us, by selling us the illusion of an easy, superficial and, of course fake 'control'.

A promising framework for meaningful human control is developed in Santoni de Sio and Hoven (2018). This new conception of control is supposed to replace the operative conception in contexts where artificial autonomous systems are causally, but not morally, responsible for outcomes. What distinguishes this account from other conceptions of control is (a) that for meaningful human control, it is not necessary nor sufficient for an agent to be able to causally intervene in a system, (b) that it is sensitive to the epistemic conditions necessary to have the kind of control to render an agent responsible, (c) that it applies to the entire 'socio-technical system' (and not just to intelligent artefacts) and (d) that it is meant to provide general design guidelines to achieve the required kind of control.

Santoni de Sio and Van den Hoven present two necessary conditions: a tracking and a tracing condition. The tracking condition requires a socio-technical system to be responsive to the relevant normative reasons to act. The tracing condition requires that one or more human agents are present in the system design history or use context who appreciate the capabilities of the system and their own responsibility for the system's behaviour. Because the tracking condition requires a system to be responsive to the reasons of the relevant agents, it is not necessary that an agent also causally influences the system's actions. So if properly implemented, this condition establishes that the decisions of the system always align with a specific agent's reasons. Moreover, the tracing condition makes sure that there are human agents who understand that they are responsible for the actions of the systems. If these conditions are met, then the system is under meaningful human control. This means that responsibility gaps can be avoided (see also Santoni de Sio & Mecacci, 2021). This conception also avoids the other conflict with our moral judgements. If an agent is merely 'in the loop', that is, is able to causally intervene, then this does not entail that she has meaningful human control over a system. The reason for this is that meaningful human control requires that a system is responsive to the agent's reasons. And, if one is able to causally intervene but does not have the knowledge required to intervene appropriately, the system will still not be responsive to the agent's reason.

3.2 Critical Thinking

3.2.1 Operative Conception

Critical thinking (CT) is a form of goal-oriented thinking (Hitchcock, 2018), a process meant to arrive at a practical decision about 'what to believe or do' (Ennis, 1962). There are mainly two ways of understanding critical thinking: as a descriptive and as a normative term. The descriptive notion of critical thinking captures the kind of thinking which is not routine and contrasts it to more automatic modes of reasoning such as a logical deduction and the application of a rule to a class. However, we will focus on the normative conception of critical thinking since we think that this conception is put under strain by technological change. The normative conception of critical thinking became popular in the latter half of the twentieth century, in the aftermath of the Second World War, when educational researchers and philosophers asked themselves: what can be done to

prevent populist dictatorships from happening again in the future (Hitchcock, 2018). Back then, it was claimed that if the majority of the population would be endowed with critical thinking skills, then anti-democratic regimes would be much harder to instate since critical citizens would see right through attempts of manipulation, propaganda, and populist discourse which are the signature moves of authoritarian regimes (Stanley, 2015). In this political context, education for critical thinking was aimed at instilling the skills and virtues that allowed students to spot bullshit (Frankfurt, 2009) and manipulation in mass-media texts or in political speeches. The societal and moral value of critical thinking was to support citizens in the exercise of democratic virtues by empowering them to avoid the pitfalls of anti-democratic discourses. Normative critical thinking is thus defined as thinking in the service of democratic goals, meaning thinking that helps individual citizens take informed actions and to evaluate existing information broadcasted to them.

The distinguishing feature of critical thinking from other kinds of thinking is its difficulty: Critical thinking is not a routine cognitive operation, such as recalling information or applying a routine procedure for solving a problem. Rather, it is deliberate and entails some cognitive effort: it 'only occurs when the reasoning, interpretation or evaluation is challenging and non-routine' (Fisher, 2019, pp. 29–30) for the epistemic agent doing this thinking. Because of this effortfulness of CT, it is a costly process to enter into, and epistemic agents need to decide for themselves under what circumstances it is worthwhile to launch into critical thinking. The operative conception of critical thinking presupposes the autonomy of the thinker, since critical thinking is usually described as thinking for oneself and refusing to delegate one's epistemic agency to authorities. From this perspective, critical thinking could be dangerous (one can only think about conspiracy theorists who also 'think for themselves' and systematically distrust public sources of epistemic authority such as scientific institutions) if taken too far, and this is safeguarded by ensuring that critical thinking is described not only as intellectual autonomy, but as a cluster of epistemic virtues such as 'curiosity, open-mindedness, attentiveness, intellectual carefulness, intellectual courage, intellectual rigour, and intellectual honesty ' (Baehr, 2013, p. 248). Since any of these intellectual virtues are laudable having in themselves, the distinctiveness of critical thinking lies in the claim that having these particular virtues *combined* will determine the epistemic agent to launch into a relentless pursuit of the truth of the issue at stake, regardless of the personal costs in terms of effort and social costs (such as becoming an outcast because one asks the uncomfortable questions) and this will lead, ultimately and by aggregating multiple critical thinking processes, to achieving the epistemic goals of a democracy: taking informed decisions primarily based on facts and on a shared interpretation of those facts. We should notice that the way the operative conception of critical thinking is described is in terms of properties of the individual epistemic agent (virtues, skills, dispositions) and assuming that one can launch in a process of critical thinking regardless of how hostile an environment is.

3.2.2 Conflict with Moral Judgements

When social media platforms gained popularity, certain behaviours emerged as deeply problematic from an epistemic perspective: Users were sharing misinformation, engaged in polarising debates while demonising the opposing side and some users even self-radicalised after watching increasingly aggressive videos suggested

by a platform's algorithm (see Alfano et al., 2018). It is reasonable to believe that some of these behaviours were taking place before social media (see gossip spreading as a model for misinformation sharing), yet the new technology of social media platforms made these behaviours highly visible and also more toxic for others. Regardless if we were to explain these behaviours as intellectual vices or as faulty cognitive heuristics (Broncano & Carter, 2021), these had rippling consequences since any user could engage in problematic epistemic behaviours and it became impossible to predict who would be affected by this. Consequently, philosophers described social media platforms as a direct threat to democracy (Sunstein, 2018) because of how easy it became to manipulate users simply by truncating the information that reached them — the Cambridge Analytica scandal of 2016 being one of the key examples.

To tackle the epistemic threats emerging from social media platforms, media scholars advanced the idea that critical thinking should be promoted more by stepping up educational interventions or by nudging users to think critically when online (Blair and (Ed.). 2019). The conception of critical thinking used here was the operative conception of critical thinking as a cluster of individual character traits that the epistemic agents acquire and then employ. However, this operative conception did not scale up to the problems posed by the new informational environments of social media. For one thing, social media platforms are an ambiguous epistemic environment where context is easily collapsed (Marwick & Boyd, d. 2011); this means that users cannot reasonably decide when they should be thinking critically about a post and when they can simply ignore it. The epistemic ambiguity of social media means that a statement posted by someone on social media can be genuinely informational for some users, while for others, it counts as noise, in an unpredictable way.

The operative conception of critical thinking gives rise to conflicting moral intuitions when applied to social media. On one hand, it has been argued that when users engage in problematic epistemic behaviours on social media (misinforming, polarising, sharing propaganda), they are in fact engaging in a kind of moral harm towards the unseen others, the users that will encounter their posts, because there is a kind of epistemic injustice that we facilitate by polluting the informational environment that others cohabitate. The operative conception of critical thinking, individualistic at its core, would have us blame the users for their presumable lack of CT or their failure to engage in critical thinking when they are online. However, there is a strong moral intuition that users are not to be blamed for being immersed in toxic informational environments since the problem is structural. To put it briefly, operative conceptions of critical thinking would commit us to 'victim blaming' by attributing exaggerated agency to the individual social media users and taking the blameworthiness of platforms out of the picture.

3.2.3 Target Conception

While the operative conception of critical thinking was describing an individual's lifelong cognitive skills aimed to fight propaganda by challenging its claims, social media platforms showed that the environment in which we act as epistemic agents matters just as much as our pre-existing skills and dispositions. With social media platforms

oriented towards personalisation of user experience, we become trapped in a filter bubble (Pariser, 2011), whereby we only see things that we agree with and we are not confronted with opposite opinions. This filter bubble is invisibly created by algorithms which deliver what they estimate we may want to see, with the purpose of making us more engaged on the platform. Social media platforms achieve an addictive effect by signalling to users that they are right in every aspect: the music they like, the activities they choose and, when they care about such things, their political views are shown as the best in the world. In this context of continuous affirmation of the self, it is very hard to deliberately put in the effort to think for oneself, to look for counter-evidence or to even become aware that one has cognitive biases. When users are immersed in social media environments where information is overwhelming them through its sheer quantity while also highly personalised, enacting the classical dispositions of critical thinking such as self-restraint, curiosity about other points of view and intellectual humility becomes increasingly difficult. Social media thus enacted a conflict between our epistemic values needed for safeguarding democracy and the conception of critical thinking as an effortful individual process. Solving the conflict would entail us committing to the idea that all social media users need to be critical about all information they see on social media, since we cannot predict which piece of information would be toxic for democracy. This is not feasible in practice, and it entails trivialising the concept of critical thinking which was designed in the first place to be a mode of thinking which was not to be deployed in everyday circumstances.

Recent work in cognitive sciences has made popular the notion of nudging (Thaler & Sunstein, 2008) which shows that we do not need to reflect consciously all the time in order to make rational choices if we can delegate to our environment some of these choices. Nudging is about designing environments which promote our values by default without us having to make these choices (think about a cafeteria where the healthy foods are placed at eye level and easier to reach) while boosting is about creating learning environments where users learn to take boost their skills for choosing in line with their values in an effortless manner by relying less on deliberate reflection and more on intuitions. Nudging and boosting show that it is possible and worthwhile to take some of the cognitive effort off the individual users and look at the design of our cognitive environments.

Social media platforms appear as cognitive environments tailored toward entertainment and confirmation of pre-existing biases and as such will not incentivise critical thinkers to apply their skills (also see Williams, 2018 on distraction by design and Voinea et al., 2020 on the cognitive detrimental effects). This does not mean that critical thinking is impossible online, but that it becomes an uphill battle that most of us do not know we are facing when we open our social media platform of choice. Furthermore, this difficulty of engaging in critical thinking cannot be attributed to user weakness of will or lack of education, but instead highlights the role that the cognitive environment plays in exercising our capacities for critical thinking, namely the key role played by dispositions. The operative conception of critical thinking appears inadequate when facing the novelty of social media as epistemic environments and needs to be revised. The target conception of critical thinking needs to accommodate the understanding that the environment plays a role in how well we think — critically or not — and this needs to be part and parcel of

the new conception of what a critical thinker should do. The target conception of critical thinking is still normative, but it designates the unity of thinker and the cognitive environment, taken together, by acknowledging that there are no free floating dispositions to be critical; rather, these are triggered by friendly environments or stifled by hostile ones. Such critically friendly environments could be designed from the beginning following certain design principles such as increasing user friction and diversifying the user's informational diet with information sources and generally avoiding to personalise the user's experience. Thus, a critical online environment would be targeted not towards maximising engagement or entertainment, but towards fostering reflection and self-knowledge in an ecosystem of human-technology interactions. There are already experiments in design showing how nudging towards online critical thinking can be achieved by changing the environment. In these experiments, users were primed to themselves if a certain piece of news was misinformation (Lutzke et al., 2019) and, after a few iterations, the users took this new habit with them onto other platforms. The new conception of critical thinking recognises that the thinker is embedded in an environment which needs to be designed for criticality in a deliberate manner, through either nudging or boosting.

3.3 Freedom

3.3.1 Operative Conception

Freedom is a fundamental concern for most normative political theories. A crucial point of reference for the philosophical debates on social and political freedom is Isaiah Berlin's distinction between a negative and a positive interpretation of freedom. The negative conception refers to the actual absence of relevant interferences or constraints on one's actions, while the positive conception concerns self-realisation and self-determination (Berlin, 1969).

More traditional political theories, such as liberalism and libertarianism, endorse a negative definition of freedom. These theories concern themselves primarily with the external sphere of action of individuals and claim that individuals should not be unduly interfered with by the State or other actors or bodies. Under this traditional conception of freedom, one's freedom is restricted when there is interference, understood as an intentional and actual intervention by other people that restricts the number and quality of the set of options or choices before agents, or as Berlin put it, that affects what doors and how many doors are open to the agents (Berlin, 1969: xlviii).

3.3.2 Conflict with Moral Judgements

Currently, we see increased concerns regarding the power of Big Tech amongst several groups in society. Scholars have increasingly devoted their attention to the effects of big data on democracy (e.g. Zuboff, 2015; Nemitz, 2018; Macnish & J. Galliott, , 2020), even wondering whether democracy will survive this new trend (e.g. Helbing et al., 2019). In addition, governmental bodies are becoming more active in controlling Big Tech (Ovide, 2021). This increased attention suggests that

traditional freedom as non-interference can be conceptually and morally deficient if confronted with the kind of power and influences that are exerted via new emerging technologies, such as machine learning or Big Data. Such technologies pose new risks that have led scholars to talk about hypernudges, which are Big Data nudges ('big nudging') that can shape people's behaviours and their choice context in a more efficacious, targeted and pervasive way through the extraction and collection of their data (Yeung, 2017).

Consider the Facebook-Cambridge Analytica scandal, where millions of people's data were unknowingly gathered and used for political profiling. The public uproar that resulted from this scandal was partly due to privacy violations and a feeling that millions of users were manipulated. The liberal conception of freedom defended by 'pure negative theorists' often implies an exclusive emphasis on interferences conceived as options-removals, or any conducts or dispositions of some other persons that prevent an action, rendering such action impossible to perform (Carter, 1999; Carter & Kramer, 2008; Kramer, 2003). However, our moral judgements about this scandal and its public consequences are not adequately exhausted by such an account of freedom. Indeed, no removals of options or conducts that prevented individuals' actions occurred prima facie. Nonetheless, such a scandal has shown a peculiar restriction on freedom that stands in need of normative justification and appraisal. There are different cases of interferences or constraints in social life — coercive or otherwise, such as manipulative ones, physical or psychological, actual or not — and the stakes of conceptually individuating and normatively justifying them are high.

In addition, filter bubbles decrease the possibility for individuals to be exposed to debates outside their own preference (Pariser, 2011). Mill (2003, specifically ch. 2) already argued for the importance of being exposed to other points of view outside your personal realm. Conflicts of opinions enhance democracy, but mostly are important to exercise one's freedom and autonomy. The traditional liberal conception does not adequately account for this since less coercive restrictions are not immediately considered freedom-restrictive, yet with the increased phenomenon of filter bubbles in combination with Mill's idea of 'conflict' exposure, this conception no longer seems to hold. However, the effects of new technological influences are not merely visible on an individual level; democratic harm is equally done on a collective level (Macnish & J. Galliott, 2020). The conceptual insufficiency of freedom as non-interference in this case is arguably motivated by the fact that such conception tends to neglect interpersonal relations and social standing and to place its predominant focus on the set of options and choices an individual enjoys.

Furthermore, what we see is that there is no satisfying possibility to check the power of big tech companies. Scholars such as political scientist Francis Fukuyama draw attention to this intuitive concern (Fukuyama et al. 2021). Regarding social media platforms, these authors point out that the real issue relates to the question of who is in charge. Twitter may suppress and fact-check Trump (now even having banned Trump), but who is suppressed depends ultimately on the person in charge, less so on the justice of the suppression. Fukuyama et al. (2021) compare the power of these companies with a loaded gun on a table: Right now, nobody picks it up to shoot, but we may wonder to what extent it is safe to leave it there. The authors

correctly refer to the necessity of checks and balances within a liberal democracy. The gun example illustrates that, currently, these companies can exercise their power without being controlled by a body who oversees their power and decisions.

This lack of checks and balances suggests a deeper concern related to interference and non-interference: is there robust non-interference? Consider again the gun example: even if these companies do not interfere with you, the fact that they can if they wish implies an unequal political relation between the companies and society in general. Indeed, the huge amount of data these companies have on a person makes it possible to single them out if they wish: you only have 'nothing to hide' if you are not explicitly on their radar. Since there are currently no adequate checks and balances mechanisms, these companies have the possibility to exercise their power unaccountably. This unaccountable exercise of power implies that freedom does not depend on interference or non-interference, as traditional liberals believe, but perhaps more so on whether there is an insurance for non-interference, in other words, whether the interference is in fact robust.

Robust non-interference depends on controlling entities. An adequate checks and balances mechanism requires not just providing control after the event; it includes providing checks on these new influences beforehand. The vast increase in technology companies simply exceeded legal regulations and public scrutiny. Only recently do these companies face increased pressures from the law, governments and the public that suggest a change in relation between these companies and society.

3.3.3 Target Conception

The intuition regarding the problems with potential interference matches a recent scholarly discussion on freedom which has been motivated by a desire to overcome the distinction between a negative and a positive interpretation of the concept. In particular, among those theorists, there are some that have identified and promoted a view of freedom as 'republican' freedom. Contemporary neo-republican political theorists such as Philip Pettit or Quentin Skinner have attempted to go beyond negative freedom, understood as the absence of interference and have individuated the conceptual core of freedom in 'non-domination' (Pettit, 1997; Skinner, 1998, 2002). Freedom as non-domination is a status, namely the enjoyment of a position that guarantees that no interference from arbitrary kinds of power is exerted (Pettit, 2011, 2012). A status like that of a slave makes him susceptible to being interfered with by a master, independently of any actual interference from the latter. Therefore, a politically worthy society is the one that maximises in its institutions and mechanisms such a conception of freedom according to neo-republicans.

Interestingly, this neo-republican framework is precisely what Fukuyama et al. seem to describe. Indeed, it is not the actual interference that worries the neo-republican; it is the potential of interference when there are no mechanisms available to hold the power accountable. Although not expressed in these terms, Fukuyama et al. describe the concept of domination regarding tech companies and society. Their possibility to interfere without having to face consequences for their actions defines domination as understood by neo-republicans: being subjected to a superior and unaccountable power. This debate sparked in the literature by neo-republicans and the situation as we encounter in today's reality suggests that the operative

conception of freedom as non-interference does not provide sufficient explanation why we should worry about the power of these companies. An existing conception of freedom such as the one endorsed by neo-republicans can be a better candidate to frame and advance the conception of freedom in the context of new emerging technologies. Indeed, it shows that political and social freedom is not about the absence of actual interference or about the doors that are open to individuals, but rather, it requires that no doorkeeper has the power to close or conceal a door without a cost (Pettit, 2011: 709). In terms of our moral intuition about facts, this means that the main concern with these companies is that they have the power to interfere with their users without being held (adequately) accountable.

4 Discussion

As we have seen, we can make progress in the philosophy of technology by engaging in some form of morally motivated conceptual engineering. In this section, we will reflect on the case studies from the previous section to see how the open questions in the conceptual engineering literature that we discussed in Sect. 2 are (or should be) addressed. More specifically, in Sect. 4.1, we show what conceptual engineering method fits best with an argumentative structure that was given in the discussion of the case studies. In Sect. 4.2, we show how this method can be used to answer the question when a new proposed conception 'changes the topic'. Finally, in Sects. 4.3 and 4.4, we discuss the question whether revisions are supposed to be global and whether the proposed conceptions have to be new, respectively.

4.1 Approaching Conceptual Engineering

There currently is no consensus on the methodology of conceptual engineering. Matti Eklund, for example, calls the question of what the proper methodology for conceptual engineering is one of the 'big questions [that] remain[s] entirely unresolved' (2015, p. 382). In Sect. 2, we indicated that there are two main approaches to conceptual engineering: a metaphysical and a pragmatic approach. Proponents of the metaphysical approach hold that we should use those concepts that fit best with metaphysical reality, i.e. those concepts that carve nature at its joints. One of the advantages of this approach is that it fits well with our intuitions about scientific and other empirical concepts.

On the pragmatic approach, when deciding which conception of a concept to employ, we should first determine what function, or purpose, this concept should perform in the context that we are discussing. Once we have determined what the function is, the best conception is that conception that fulfills this function best. Sometimes, the function that a concept ought to perform is the function that it has at the moment. Suppose for example that one had in one's society a conception of marriage that precludes same-sex couples. Moreover, suppose, as is plausible, that the function of marriage is to afford a special legal and social status to a range of close relationships (Cappelen, 2018). In these circumstances, we can come to see a conception of marriage which includes same-sex couples as better than the old one.

From the examples that have been discussed in the previous section, we see that the pragmatic approach is favoured. The conceptions that are the target of the engineering work all relate, or ought to relate, to a specifically moral concept, for example, when discussing which conception of *control* to use, we first pointed out that we are interested in a specific function of that concept, i.e. its relation to responsibility. Using this heuristic, we also addressed the question which conception is best. To answer this question, we looked at an example in the context; a person behind the wheel of a self-driving car that has adaptive cruise control switched on. Given some additional details, we can now truly say that the driver has causal control over the car, but lacks meaningful human control. So we cannot say that one description is true while the other is not. When deciding which conception is best in the context of autonomous intelligent systems, we based our decision on the question which conception relates best to appropriate responsibility attribution. The reason for this is that this is the conception that fulfils the context-determined function best.

The same can be said for *freedom* and *critical thinking*. When we assess which conception of *critical thinking* is best, we have a specific goal in mind: making sure that the kinds of dictatorships that we have seen in the past will not happen again, hence that democratic values are supported by informed citizens. Given this role, it was argued that a target conception, which includes critical environments, was able to fulfil the normative constraints that the operative conception could not. In Sect. 3.3, about *freedom*, the function that was highlighted was to indicate a desirable relation to those that are powerful. It was argued that in the context of influential social media companies, the neorepublican conception of freedom does a better job indicating which relation is desirable than the negative conception of freedom. Whether the conception is indeed better in fulfilling a specific function is a matter of normative debate. And this is of course exactly what the pragmatic approach predicts (Thomasson, 2020, pp. 451–455).

4.2 Are We Changing the Topic?

Some philosophers are critical of the practice of conceptual engineering in general, because they think that concept(ion) revision is always just changing the topic. The risk is that '[r]evisionary projects are […] providing answers to questions that weren't being asked' (Haslanger, 2012, p. 225). Indeed, sometimes, conceptual engineering really does change the topic. An example of this approach is the introduction of the concept of online harassment that questions our traditional understandings of moral wrongs, intentionality and evil (Cocking & Hoven 2018). A different example is Miranda Fricker's work on 'epistemic injustice' (Fricker 2007). Fricker recognized that we lacked the conceptual tools to represent a certain kind of injustice: the injustice which consists in a wrong done to someone specifically in their capacity as a knower. Introducing this new concept made it possible to discuss different instances of this injustice under one heading. But we do not always need to introduce a new concept. Sometimes, we remain on the same topic (i.e. concept) and propose a new answer to an old question, for example, this was the case for critical thinking that, as we proposed, needs to be replaced with an enlarged conception of what counts as the subject of critical

thinking — i.e. the agent + the epistemic environment. These scenarios generate a new challenge: *How do we distinguish between scenarios where we change the topic and scenarios where we do not?*

Let us look at one of the case studies above to answer this question. Proponents of meaningful human control are also sometimes accused of changing the topic. What their critics maintain is that they are no longer talking about *control*, but about something else. The underlying accusation is that what the MHC proponents discuss is not relevant to the original discussion. And indeed, if we are primarily interested in our current conception of *control*, then an analysis of meaningful human control might not be relevant. A better way to understand the argument of proponents of MHC is twofold. First, they argue that in the current context we should be interested in those conceptions that are suitably related to attributions of responsibility. And secondly, they argue that with regard to autonomous intelligent systems, MHC performs this function best. Remember that we stated in the introduction that the question of what counts as remaining on topic depends on context-sensitive similarity relations. The first argument is supposed to fix the similarity relation and thereby the set of conceptions that are 'on topic'. Given this set of conceptions, the second argument now indicates which of these conceptions is best.

The same strategy can be employed for the second and third case study. The enlarged conception of *critical thinking* and the re-engineered conception of *freedom* both play the same functional role: Conceptions of *freedom* indicate a desirable relation to those that are powerful and conceptions of critical thinking indicate ways of thinking which make sure that democratic values are supported by informed citizens. In these cases, we can therefore say that we stay 'on topic' when we introduce a different conception, because that conception plays the same role as the operative one.[10]

4.3 Do We Aim for a Purpose- or Context-Specific Revision, or Is the Revision Supposed to Be Global?

In Sect. 2, we have presented two views on conceptions that are sometimes defended in the literature: conception relativism and conception imperialism. The conception relativist holds that no coherent conception is better than another one and the conception imperialist maintains that only one conception of a concept is correct in all contexts. This brings us to the question: *Do we aim to revise a concept for a particular context (i.e. locally) or rather generally and globally?* We see that in all projects we discussed, conception relativism was implicitly rejected. A conception can be completely coherent and still be deficient in different ways. All examples made clear that conceptions can be morally deficient in specific contexts. We have, for example, shown that traditional common-sense and engineering conceptions of control should be problematized in the light of developments in digital technologies.

[10] We do not want to commit ourselves to the claim that playing the same functional role always entails sameness of topic (as Haslanger 2012 seems to suggest; but see Riggs 2021 for criticism of this idea). What we claim is that playing the same functional role is sometimes a reason that two conceptions are 'on topic'. This leaves open the possibility that in some contexts other factors tell us which conceptions are legitimate revisions.

These examples also show, however, that conception imperialism is at least sometimes rejected. If conception imperialism is correct, then revisions are always global, that is, if a conception is best in one context then it is ipso facto the best in all contexts. *Meaningful human control* is, for example, a conception of *control* that is explicitly supposed to be local. The tracking and the tracing conditions refer to design-histories and socio-technical systems and are therefore not suited for many other contexts (e.g. when discussing the question whether a human has control over her own bodily movements). A conception imperialist who wants to defend this conception of meaningful human control would therefore have to deny that 'regular' control and 'meaningful human control' deal with the same topic. But as we have seen above, this is highly implausible given the purposes that we have.

Concerning *critical thinking,* both the operative and the target conceptions are context-dependent. The normative context of the operative conception of *critical thinking* is wherever a democratic process is called for. The problem that *critical thinking* has to solve remains, both for the operative and the target conceptions, what to do when decision-making actors are confronted with conflicting evidence, when information and misinformation compete and one has to think for oneself which one to trust. This problem appears not onyl in most processes that strive to be democratic, be it at a political level, but also in teams of co-workers, in any process where multiple actors relate horizontally to each other and need to arrive at a decision together. The context of social media platforms is a specific form of decision-making, whereby one needs to decide for oneself what to believe and do, regardless if the decision is political or not. Hence, we are not changing the topic by changing the context, since social media platforms are specific instances of horizontal decision-making in asymmetrical flows of information, whereby the architecture of the information flows is technology-dependent.

From the limited number of cases that we present in this paper, there is only one conception that can arguably be said to be a global conceptual revision: *freedom*. The neo-republican conception of freedom has been applied in other contexts as well,[11] and a case can be made that this conception is the only correct one in all contexts. What is important, however, is that one is not committed to this position. It is even possible, and not even far-fetched, to reject Skinner's and Pettit's proposal for the neo-republican conception of freedom in the domains that they are interested in and hold that this conception of freedom is superior in the context of digital technologies. It is important to note that maintaining that the neo-republican conception of freedom is appropriate for all contexts, in which *freedom* is used, does not commit one to conception imperialism. The claim that for a specific concept there is only one appropriate conception is compatible with the claim that there are other concepts that have different appropriate conceptions in different contexts.

4.4 Is There Already a Candidate Conception Available, or Should We Construct a New Conception?

The fourth issue that is relevant is whether the conception that is proposed is new, or if a conception that is used in a different context can be used for the new context as well. In that case, we might consider a variety of contexts, e.g. disciplinary context, application

[11] For example, example medical care (O'Shea 2018 and immigration (Costa 2016).

context or historical context. John Rawls, for example, argued for justice as fairness in the context of liberal societies. The aim of his revision was narrow, i.e. only supposed to apply to a specific context. However, this does not rule out the possibility of applying this conception of justice appropriately in a completely different context. As we have seen in Sect. 3.3, for example, the neo-republican conception of freedom can be used in contexts that were not envisioned when this conception was introduced. On other occasions, the conception we want to propose is an entirely new one (e.g. the original introduction of justice as fairness). The fourth issue is therefore: *Is there already a candidate conception available elsewhere, or should we construct a new conception?*

As Sect. 3.3 has shown, reviving older or less popular theories help us address some of the moral judgements we have with the new technological influences and society. Originally, freedom from domination has been used in particular to address institutional arrangements, like the relation between the State and her citizens. This conception of *freedom*, however, may be used in new settings, such as the relation between companies and their users.

In Fig. 1, we illustrate this question in an (oversimplified) diagram: One concept (freedom) includes several conceptions (e.g. freedom as non-interference and freedom as non-domination) that correspond to a particular context. Where initially the conception 'freedom as non-interference' would relate to the context 'social media', Sect. 3.3 illustrates how this traditional conception conflicts with our moral judgements. The question then is whether we require a new conception of freedom or an already existing alternative. As argued in Sect. 3.3, the already existing alternative 'freedom as non-domination' is a good fit in this context and hence there is no need for developing a new conception of freedom to accommodate our moral judgement. The red cross in Fig. 1 reflects our conflict and the green arrow our moral judgement. For the concept of freedom in the context of social media, the answer to the question whether we should construct (1) a new conception or (2) is there already a candidate conception available elsewhere therefore concludes: there is already a good alternative available, namely freedom as non-domination.

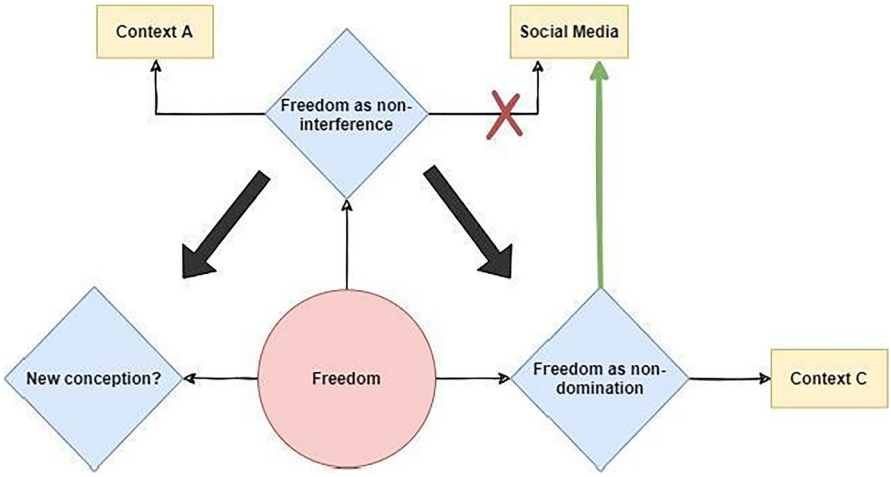

Fig. 1 Schematic illustration of question 4 applied to the concept of freedom in the context of social media

We see that by expanding this already existing conception to a new realm — namely, ethics and philosophy of technology, we can better formulate what needs to be done to address concerns with these new technologies. For instance, for neo-republicans, it is important that citizens have the opportunity to *contest* governmental decisions, as this provides a checks and balances system (Pettit, 1997, 2012). With Facebook-Cambridge Analytica, people were not even aware that their data was being used for political profiling, which makes contestation rather difficult.

Not only does the new conception of freedom as non-domination meet some moral judgments with the power of emerging technologies — specifically in the field of Big Data — but it also provides a way to formulate policy and legal regulations that are necessary to ensure users' freedom and autonomy. Freedom from domination hence proves to be a promising candidate for managing issues arising in the Digital Age.

The case study of *meaningful human control* illustrates a similar point. On a purely causal conception of *control*, someone in the driver seat of a semi-autonomous vehicle who is physically capable of intervening in the car's operation is thereby morally responsible. Moreover, on this conception, fully autonomous vehicles are morally problematic because they seem to introduce responsibility gaps. To remedy this problem, Santoni de Sio and Van den Hoven (2018) have adopted a conception of control from a different context: guidance control (Fischer & Ravizza, 1998). They subsequently modified the conception to make it suitable in the context of autonomous artificial agents. So this is another example in which an existing conception was introduced in a novel context.

5 Conclusion

In this paper, we have presented three case studies that can be interpreted as concrete examples of conceptual engineering. In all three cases, we identified an operative conception that we argued was morally defective. Certain conceptions fit better with our moral judgements than other conceptions, and, as we have seen, this moral defect is an important, albeit defeasible, reason to engage in conceptual engineering. We have argued that this can be understood as an instance of the pragmatic approach to conceptual engineering. The moral defect is a reason for conceptual change, because it is part of the function of that concept to contribute to that specific moral value.

We showed that the prime reason for opting for a different conception of *control* is that the new conception has a better relation to responsibility attributions. For the concept *critical thinking*, we showed that the concept ought to promote the support of democratic values by encouraging citizens to take responsibility for evaluating information. Consequently, we argued that the new conception fulfils this function better than the old one. For *freedom*, we argued that the neo-republican perspective better captures morally problematic power relations than the more traditional freedom as non-interference.

We have also shown that the moral adequacy of our conceptions is context-dependent. Conceptions that are morally adequate in existing contexts can be shown to have moral defects in new contexts. This is why the disruptive nature of new technologies functions as an important driver for work in the ethics of technology. When

new socio-technological ecosystems are introduced, new contexts in which our operative conceptions are evaluated are introduced as well.

In the final section, we aimed to make explicit what answers engineering philosophers implicitly give to a number of open methodological questions concerning conceptual engineering. As such, this paper makes two contributions to the current literature. Firstly, it contributes to the literature on conceptual engineering by presenting three cases in which conceptual revisionary work has actually been done and has direct real world consequences. Secondly, we believe that it would be useful if future conceptual research were more integrated and explicitly interwoven with existing methodologies in the philosophy of technology (such as values sensitive design and design for values).[12] Our hope is that our findings are helpful for such a project.

Acknowledgements We would like to thank Jeroen van den Hoven, Filippo Santoni de Sio, Georgy Ishmaev and all other members of the Delft Digital Philosophy Seminar for helpful feedback on an early draft of the paper.

Author Contribution HV developed the concept of the paper and took the lead in writing it. LM wrote the section on critical thinking, JM and MC wrote the section on freedom. All authors were involved in revisions of the paper. The schematic representation was devised by JM.

Funding The research of Jonne Maas reported in this work was partially supported by the EU H2020 ICT48 project 'Humane AI Net' under contract no. 952026. Lavinia Marin's work in this project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 707404. This work is part of the research programme Ethics of Socially Disruptive Technologies, which is funded through the Gravitation programme of the Dutch Ministry of Education, Culture and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031). The research by Jonne Maas, Lavinia Marin and Herman Veluwenkamp is also supported by the Delft Digital Ethics Centre.

Declarations

Ethics Approval This is a theoretical study. No ethical approval is required.

Consent to Participate This is a theoretical study. No consent is required.

Consent for Publication This is a theoretical study. No consent is required.

Competing Interests The authors declare no competing interests.

Disclaimer The opinions expressed in this document reflect only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

[12] Value sensitive design is a design process which involves three types of investigation (Friedman & Hendry 2019). In one of those investigations, the conceptual investigation, there already is attention to the conceptions used by the different stakeholders. We take it that our work can help improve this type of investigation.

References

Alfano, M., Carter, J. A., & Cheong, M. (2018). Technological seduction and self-radicalization. *Journal of the American Philosophical Association, 4*(3), 298–322.

Anderson, E. (2015). Equality and freedom in the workplace: Recovering republican insights. *Social Philosophy and Policy, 31*(2), 48–69.

Baehr, J. (2013). Educating for intellectual virtues: From theory to practice. *Journal of Philosophy of Education, 47*(2), 248–262. https://doi.org/10.1111/1467-9752.12023

Carter, I. (1999). *A measure of freedom*. Oxford University Press.

Carter, I., & Kramer, M. H. (2008). How changes in one's preferences can affect one's freedom (and how they cannot): A reply to Dowding and van Hees. *Economics and Philosophy, 24*(1), 81–96.

Cocking, D., & Van den Hoven, J. (2018). *Evil online*. John Wiley & Sons.

Costa, M. V. (2016). Republican liberty and border controls. *Critical Review of International Social and Political Philosophy, 19*(4), 400–415.

Deutsch, M. (2020). Speaker's reference, stipulation, and a dilemma for conceptual engineers. *Philosophical Studies, 177*, 3935–3957.

Eklund, M. (2012). Alternative normative concepts. *Analytic*. *Philosophy, 53*(2), 139–157.

Fisher, A. (2019). What critical thinking is. In A. Blair (Ed.), *Studies in critical thinking* (pp. 7–32). University of Windsor.

Frankfurt, H. G. (2009). *On bullshit*. Princeton University Press.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be? *Noûs, 34*(1), 31–55.

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

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Van den Hoven, J., Zicari, R. V., & Zwitter, A. (2019). Will democracy survive big data and artificial intelligence? *Towards digital enlightenment* (pp. 73–98). Springer.

Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice, 22*(3), 731–747. https://doi.org/10.1007/s10677-019-10007-9

Kramer, M. H. (2003). *The quality of freedom*. Oxford University Press.

Leavitt, S. (2006). 'A private little revolution': The home pregnancy test in American culture. *Bulletin of the History of Medicine, 80*(2), 317–345. https://doi.org/10.1353/bhm.2006.0064

Löhr, G. (2021). Commitment engineering: Conceptual engineering without representations. *Synthese, 199*, 13035–13052. https://doi.org/10.1007/s11229-021-03365-4

Macnish & J. Galliott, (Ed.). (2020). *Big data and democracy*. Edinburgh University Press.

Marwick, A. E., & Boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1), 114–133.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mill, J. S. (2003). *On liberty*. Yale University Press.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180089.

Nyholm, S. (2018). The ethics of crashes with self-driving cars: A roadmap I. *Philosophy Compass, 13*(7), e12507. https://doi.org/10.1111/phc3.12507

O'Shea, T. (2018). Disability and domination: Lessons from republican political philosophy. *Journal of Applied Philosophy, 35*(1), 133–148.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin.

Pettit, P. (1997). *Republicanism: A theory of freedom and government*. Oxford University Press.

Pettit, P. (2011). The instability of freedom as non-interference. *The Case of Isaiah Berlin Ethics, 121*, 693–716.
Pettit, P. (2012). *On the people's terms*. Cambridge University Press.
Rawls, J. (1999). *A theory of justice*. Oxford University Press.
Riggs, J. (2021). Deflating the functional turn in conceptual engineering. *Synthese, 199*(3–4), 11555–11586.
Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy and Technology, 34*(4), 1057–1084.
Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers Robotics AI, 5*, 15.
Skinner, Q. (1998). *Liberty before Liberalism*. Cambridge University Press.
Skinner, Q. (2002). A third concept of liberty. *Proceedings of the British Academy, 117*(237), 237–268.
Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x
Stanley, J. (2015). *How propaganda works*. Princeton University Press.
Sunstein, C. R. (2018). Is social media good or bad for democracy. *SUR-Int'l J. on Hum Rts., 27*, 83.
Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines, 30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4
Van den Hoven, J. (2013). *Value sensitive design and responsible innovation* (pp. 75–83). Managing the responsible emergence of science and innovation in society.
Voinea, C., Vică, C., Mihailov, E., & Savulescu, J. (2020). *The Internet as cognitive enhancement*. Advance online publication. https://doi.org/10.1007/s11948-020-00210-8
Williams, J. (2018). *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge University Press.
Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society, 20*(1), 118–136.
Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology, 30*(1), 75–89.
Baker, W. D. W. P. of P. R. (2019). *The structure of moral revolutions: Studies of changes in the morality of abortion, death, and the bioethics revolution* .
Berlin, I.(1969). *Four essays on liberty*. London: Oxford University Press. New ed. in Berlin 2002.
Blair, A. (Ed.). (2019). Studies in critical thinking. University of Windsor. https://doi.org/10.22329/wsia.08.2019.
Broncano, F., & Carter, J. A. (2021). The philosophy of group polarization: Epistemology, metaphysics, psychology / Fernando Broncano-Berrocal, J. Adam Carter (1st). Routledge studies in epistemology. Routledge.
Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford University Press.
Chalmers, D.J. (2020). What is conceptual engineering and what should it be?, *Inquiry*. https://doi.org/10.1080/0020174X.2020.1817141
Ennis, R. H. (1962). A concept of critical thinking. Harvard Educational Review. 32(1): 81–111
——, .(2015). Intuitions, conceptual engineering, and conceptual fixed points. In *The Palgrave handbook of philosophical methods* (pp. 363–385). Palgrave Macmillan, London.
Eklund M (2021) Conceptual Engineering in Philosophy. In Justin Khoo & Rachel Sterken (eds.), The Routledge Handbook of Social and Political Philosophy of Language.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.
Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*, Mit Press.
Fukuyama, F., Richman, B., & Goel, A. (2021). How to save democracy from technology. Retrieved March 04, 2021, from https://www.foreignaffairs.com/articles/united-states/2020-11-24/fukuyama-how-save-democracy-technology?utm_medium=email_notifications&utm_source=reg_confirmation&utm_campaign=reg_guestpass
——, (2012). *Resisting reality: Social construction and social critique* . OUP USA.
——, (2020). How not to change the subject. In T. Marques & A. Wikforss (Eds.), Shifting concepts: The philosophy and psychology of conceptual variability. Oxford University Press.

Hitchcock, D. (2018). Critical thinking. In Edward N. Zalta (Ed.), The Stanford encyclopedia of philosophy (2018th ed.). Metaphysics Research Lab, Stanford University.

Lalumera, E. (2014). On the explanatory value of the concept-conception distinction. *Rivista italiana di filosofia del linguaggio.*

Lutzke, L., Drummond, C., Slovic, P., & Árvai, J. (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change, 58*. https://doi.org/10.1016/j.gloenvcha.2019.101964

Orlowski, J. (2020), *The social dilemma*, Exposure Labs, Netflix, netflix.com/title/81254224.

Ovide, S. (2021). The state house versus big tech. Retrieved March 05, 2021, from https://www.nytimes.com/2021/02/16/technology/the-state-house-versus-big-tech.html

Nickel, P. J., Kudina, O., & Poel, I. van de. (2020). *Moral uncertainty in technomoral change: Bridging the explanatory gap* .

Rudy-Hiller, F. (2018). The epistemic condition for moral responsibility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/

Thomasson, A. (2020). A pragmatic method for normative conceptual work. In *Conceptual engineering and conceptual ethics*, 435–457.

——, (2017). The design turn in applied ethics. In J. Van den Hoven, S. Miller, & T. Pogge (Eds.), *Designing in ethics* (pp. 11–31). Cambridge University Press. https://doi.org/10.1017/9780511844317.002

author references hidden for review purposes)