AI and Ethics
https://doi.org/10.1007/s43681-023-00295-4

ORIGINAL RESEARCH

If conceptual engineering is a new method in the ethics of AI, what method is it exactly?

Guido Löhr[1]

Received: 15 March 2023 / Accepted: 3 May 2023

Abstract
Can a machine be a person? Can a robot think, be our friend or colleague? These familiar questions in the ethics of AI have recently become much more urgent than many philosophers anticipated. However, they also seem as intractable as ever. For this reason, several philosophers of AI have recently turned their attention to an arguably new method: *conceptual engineering*. The idea is to stop searching for the real essence of *friendship* or our ordinary concept of the person. Instead, ethicists of AI should engineer concepts of friend or person we *should* apply. But what exactly is this method? There is currently no consensus on what the target object of conceptual engineers is or should be. In this paper, I reject a number of popular options and then argue for a pragmatist way of thinking about the target object of conceptual engineering in the ethics of AI. I conclude that in this pragmatist picture, conceptual engineering is probably what we have been doing all along. So, is it all just hype? No, the idea that the ethics of AI has been dominated by conceptual engineers all along constitutes an important meta-philosophical insight. We can build on this insight to develop a more rigorous and thorough methodology in the ethics of AI.

Keywords Conceptual engineering · Ethics of AI · Artificial intelligence · Conceptual ethics · Pragmatism · Representationalism · Inferential role semantics

> Our interest in philosophical concepts like PERSON and VOLUNTARY is not just to parse the world in such-and-such a way. Rather, we think that persons should be treated differently than non-persons (only they get rights…).
>
> Jonathan Weinberg [47, p. 32]

1 Introduction

Can a machine be a person? Can a robot think, be our friend or a genuine colleague? What can it mean for artificial minds to have rights? Can we make sense of the idea that a robot might have free will and act voluntarily? These familiar questions in the philosophy and ethics of AI have recently become more urgent than many philosophers anticipated. Especially in the last two or three years, immense progress in AI research—especially progress in so-called large language models (LLMs)—has generated a lot of interest in questions or problems that have already occupied philosophers and ethicists for many decades if not centuries. What seems to be new is that these rather abstract topics in the philosophy of mind are no longer merely theoretical fantasies or thought experiments. Many scholars and engineers seem to believe that in the next ten years, the question of whether a robot can be a person with rights, or whether a sophisticated deep learning application can be my friend or colleague will become of real practical relevance (see for example [12, 32] or [31] for a review).

Considering the immense increase in funding for the philosophy of AI in the past years, it seems that stakeholders from the sciences, industry, and politics are counting on philosophers to contribute to difficult philosophical questions regarding the nature and classification of AI. Alas, it seems that progress in AI research exceeds the progress made in philosophy. Difficult metaphysical questions, such as whether a robot can have rights or whether they can be sentient seem as intractable as ever. It is unlikely that we will see much groundbreaking philosophical progress on the essence of personhood or the nature of sentience in the next

✉ Guido Löhr
loehrg@icloud.com

[1] Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands

Published online: 16 May 2023

ten or twenty years. So, are stakeholders wrong to put their faith in philosophers? Are we bound to disappoint them? With or without philosophers, the problem of whether a robot can be a person, friend, or colleague will eventually be "answered" by society. The question is whether such legal or political decisions are based on good reasons and whether philosophers will sit at the metaphysical (political) tables where such decisions are made.

Because difficult metaphysical questions about AI minds seem as intractable as ever but practical problems are approaching fast, some philosophers of AI and technology like Floridi [9], Cappelen and Dever [6], Veluwenkamp et al. [45], Babushkina [1],Himmelreich and Köhler [15] or Veluwenkamp and Van den Hoven [46] have recently turned to a supposedly new methodology: the booming conceptual engineering and conceptual ethics literature (e.g., [5, 30], see Koch et al. [23], forthcoming for a review). This method, they hope, may help them avoid certain old metaphysical disputes. Conceptual engineering can be understood (in a broad sense) as the practice of intentionally designing concepts (cf., [5, 8, 18]). The aim is no longer to investigate the "real" essence of *person, right,* or *friendship* (a kind of realist metaphysical investigation, see [42] for discussion) or to describe *our* current concept of them (conceptual analysis, see [19]. The aim of conceptual ethicists and engineers working on topics of AI is to design concepts of, say, *friend* or *mind* that we *should* use, i.e., concepts that play a certain function, which is in accordance with our overall or certain more specific, perhaps epistemic, interests and values.

But if conceptual engineering is really a new method for the ethics of AI, what method is it exactly? What method *should* it be such that it can be useful to the relevant stakeholders? These questions arise from the fact that the nature of conceptual engineering is controversial. Even, or especially, the question of the target object of conceptual engineers—what it is that conceptual engineers aim to modify, has been debated [18, 23]. In the next section, I summarize the most popular theories of the target object of conceptual engineering. In Sect. 3, I reject them as unfit for the ethics of AI (they might still be permissible for other domains). In Sect. 4, I summarize a recently developed pragmatist alternative way of thinking about conceptual engineering. In Sect. 5, I conclude that conceptual engineering—construed in a more pragmatist fashion—is probably what we have been doing all along and that its limits are very familiar ethical ones. This conclusion is not a negative one, however. I argue that this insight is an important meta-philosophical one that will help us improve ethical reflection about AI. It is a chance to reflect on and develop a more rigorous value-sensitive and empirically grounded methodology in the ethics of AI.

2 Conceptual engineering as modifying application conditions

To avoid confusion, I distinguish here three practices that are usually conflated under the same label "conceptual engineering" but that should, for this paper, be kept apart. "Conceptual engineering" is what I will call the *practice* of proposing the introduction, revision, or elimination of whatever we eventually consider to be the target object of this practice (e.g., word's meanings, concepts, entitlements, norms, etc.). "Conceptual ethics" is what I call the practice of assessing these proposals developed by the engineer. Conceptual ethicists answer or investigate the question of which concepts/words/entitlements/norms we *should* choose all things considered. This assessment need not necessarily be constrained only by ethical considerations. Other considerations may be prudential or epistemic in nature [41]. Conceptual ethicists also are the ones that request certain engineering solutions from the conceptual engineer and who determine whether it is permissible or required to push or lobby for the implementation of the positively assessed concept. Finally, what I call "conceptual activism" is the practice of trying to implement or preserve new or established concepts.[1]

Paradigmatic examples of conceptual engineering projects so far have mainly involved social, scientific, and philosophical concepts like *truth* [39], *misogyny* [29], *woman* [13], the official concept of *planet* (The International Astronomical Union), or even the concept of *concept* itself [27]. All these examples have in common that they consider the conceptual status quo to be insufficient or deficient and thus essentially propose or "engineer" a new set of conditions of applications for the same or related terms (in the form of a definition or distinction) in the hope of thereby improving our conceptual tools [23]. For example, according to Haslanger [13] pg. 230, the concept of *woman* that we should use in certain discourses is the following:

[1] The same person can embody all three roles. For example, a conceptual engineer might propose to change the definition of a word (engage in conceptual engineering), but then also assess whether this definition is better than the more established one. Finally, she might argue for implementing this definition in the community by giving talks and publishing her work. Thereby, she is acting as a conceptual activist. For the purpose of this paper, these distinctions are quite helpful as we will see. Again, this distinction is more or less idiosyncratically chosen for this paper. It is not however alien to the literature and implicit in much of it (e.g., [30]). For example, it allows us to evaluate the practice of conceptual engineering independently of the probability of activists being able to implement the new concepts in the community. We can, for example, design or engineer a concept without this concept being part of the linguistic community. It is interesting what concepts we should be using independently of the question of whether it is likely that we will use them.

> S *is a woman* iff$_{df}$ S is systematically subordinated along some dimension (economic, political, legal, social, etc.), and S is "marked" as a target for this treatment by observed or imagined bodily features presumed to be evidence of a female's biological role in reproduction.

Haslanger hopes that her definition of woman helps philosophers or other interested individuals to think about the world in a way that increases their awareness of certain inequalities they would otherwise not be aware of (oppression based on gender). In the case of the word *truth*, Scharp's distinction of different concepts of truth ought to help us think about the world more coherently, e.g., by avoiding certain paradoxes. Machery (2009) proposes a set of application conditions for the concept of concept to be in psychology—a body of information retrieved by default in a stable and context-insensitive manner—only to then conclude that this concept does not refer to any natural kind and should be eliminated from scientific psychology. Machery's goal, however, is the same as Haslanger and Scharp's i.e., to improve the representational systems of certain communities. In the case of Machery, it is the representational system used in scientific psychology. Haslanger aims to improve the vocabulary of feminist philosophers. Scharp wants to improve the vocabulary of epistemologists.

How can we best describe or make sense of this practice of conceptual engineering? What is its so-called "target object", and what are conceptual engineers doing with it (cf., [18]? Unsurprisingly, this is a matter of controversy. More surprising, perhaps, is that this question does not seem to depend much on what the respective author takes concepts to be. Most members of the debate allow for the possibility of conceptual engineering being a misnomer (but see [17] for insisting that the name is a constraint). Whether the target objects of conceptual engineers are concepts is more debated than the question of what nature of concepts conceptual engineers should assume. So, what then is the target object of a conceptual engineer if it is not necessarily a concept? Or could it be more than one (see [38] or [7] for a pluralist option)? For the ethicist of AI interested in conceptual engineering, this fundamental methodological question is of course of the utmost importance given that it seems odd to engineer something that one knows very little about. In the worst case, we will end up engineering the wrong thing. For example, if we want to engineer a concept of friendship that we should use in response to social robots, then we should know whether we should engineer the *concept* of friendship or something else.

Several different target objects have been proposed, most of which can be understood as some form of what Cappelen [5] calls a "representational device". There are several different versions of the representational approach to conceptual engineering. According to Cappelen [5], the target object of conceptual engineering is the intension and extension of a word. How can the intensions of expressions be changed? Cappelen believes that this does not just entail proposing a set of criteria of application (or a definition), although it may be a significant step toward it. Instead, he has strong externalist meta-semantic commitments. According to semantic externalism, the external environment of speakers partly or to a large part determines the extensions and intensions of words. The relevant elements of the external environment include experts in the community, the history of use going back to the introduction of a term, complex patterns of use over time in the community, and what the world happens to be like (independently of what the speakers believe the world is like). But does this mean that an engineer has to change the world to change the meaning of our words? According to Cappelen, this is indeed the case. This is the reason he is skeptical about the success of many conceptual engineering proposals.

According to Mark Pinder [33], while a conceptual engineer may not be able to change the standing meaning of communities, she can still design conditions of application of words (intensions and extensions) for special local contexts, say a paper or book. Thus, according to Pinder, the fundamental target objects of conceptual engineers are not linguistic standing meanings but so-called "speaker meanings", which he understands in broadly Gricean terms. A speaker meaning is what a speaker intends to convey using a given utterance in a certain context [11]. As Grice has argued, speaker meanings and semantic meanings can be different, e.g., when the speaker is ignorant about the semantic meanings of their utterances or when they decide to deviate from semantic meanings. Thus, we can engage in conceptual engineering concerning a word by making explicit, e.g., by means of a definition, that we will speaker-mean m' by a word in certain contexts and by presenting m' as a good concept to speaker-mean by X within this context. The paradigmatic examples above seem to do exactly that: propose new intensions of words that may be advantageous for certain purposes.

According to the third class of views on the target object of conceptual engineering, defended by Edouard Machery [28], Joey Pollock [34], or Isaac [17], the target objects are what especially many psychologists consider to be concepts, i.e., certain bodies of information (called prototypes, exemplars or theories). We rely on these bodies of information in higher cognitive tasks like categorization or drawing certain default inferences from the use of a word. Changing or updating such bodies of information or inferences that ordinary speakers are disposed to draw is clearly possible. We can, for example, simply convince others of certain facts or inconsistencies in their thinking. This is not only rational (if our beliefs are false), but also what many conceptual

engineering projects can be interpreted as doing, i.e., showing why, e.g., our concept of innateness or even our concept of concept is flawed (see [28] for such an engineering attempt). Similarly, we can interpret Haslanger's engineering of the concept of 'woman' as an attempt to change people's body of beliefs about what women are or can be. One way we can change these target objects is simply by convincing others of their false beliefs.[2]

What all these accounts have in common is that they understand the target object of conceptual engineering to be the criteria for the application of representational devices (words, concepts). This seems to be a promising approach given that the paradigmatic exemplars of conceptual engineers seem to engage exactly in this practice, namely of introducing definitions or distinctions, i.e., different conditions under which a word or concept should be applied at least in a certain context. Applied to the ethics of AI, we should then primarily design new representations in the form of mainly proposing new intensions of concepts or standing word or speaker meanings—or by designing bodies of beliefs or information in the form of psychological prototypes for example. A conceptual engineer of personhood and AI will then design new intensions or bodies of information that, e.g., will include robots and that will then be assessed as to whether it is ethically or prudentially advantageous, permissible, or even required. A conceptual engineer of AI who is interested in whether robots can be our friends may design a set of criteria of application of the word 'friend' that includes robots and justifies it by certain arguments.

3 Conceptual engineering can't just be about application criteria

While a representational approach to conceptual engineering may be descriptively accurate and also reasonable to do for *some* projects, I argue that it is not what conceptual engineers *should* be doing in the ethics of AI. I argue that even if it captures what some philosophers of AI are doing (i.e., proposing criteria of application),[3] it, at best, captures only half of what they *ought* to be doing. The problem is that ethicists of AI cannot *just* be interested in the application conditions of words or concepts and the question of whether they can best fulfill a certain function. They cannot be interested in just proposing intensions and extensions and they cannot just be after changing people's inferential dispositions. Ethicists of AI must *also* be interested in the normative or ethical consequences of an application of a word or concept. The reasons are that, first, designing intensions and extensions alone or dispositions to apply a certain word to an object is impossible to ethically assess without also determining the consequences of applying them in the proposed way. Second, but relatedly, and perhaps more urgently, it would also give doing conceptual engineering in the ethics of AI what Jorem and Löhr [20] call a "bad rationale". Let me explain:

Expanding the extension of a concept by reducing the intension (e.g., to include robots) is often initially motivated by reasons pertaining to consistency [39] or because it helps us generate a representation that serves a certain function, e.g., to make sense of the output of robots or render them responsible for certain events [15]. For example, if what we really value about the concept of a person or friend can also apply to robots, then we might not have a good reason to exclude robots from such applications except a kind of anthropocentrism. Consequently, it might make sense to engineer the intension of the concept of friendship such that it includes certain robots. Such an approach tacitly assumes that the broader consequences of application remain the same. Whether this should be the case, however, should be part of any conceptual assessment and in some cases may, however, be difficult to maintain. For example, if we open up the concept of friendship to certain sophisticated robots, does this mean that the robot is entitled to loyalty, which we normally expect a friend to have? If we open up the concept of colleague to robots, does this mean that a robot should be considered for promotion in the same way other colleagues are? Both of these latter inferences seem to be difficult to even make sense of unless the robot were the kind of system that could reasonably be loyal or benefit from a promotion.

Therefore, the inferential consequences of opening the application conditions of words and concepts up to new kinds have to be part of an engineering proposal (the design process) in the ethics of AI. Otherwise, there is no reason to propose a change at all. Intensions and extensions are

[2] According to Koch's [22] and Sawyer's [38] hybrid views, conceptual engineers should not only engineer dispositions to make certain inferences but also the standing meaning of words and concepts. Their reason is that if we change people's applications of words, we might end up using words incorrectly. Koch even argues that concepts have two components or contents (a dual content view). One part of a concept's content is the "referential content" that Cappelen and Pinder want to engineer and that is understood as intensions of concepts, words, or speaker meanings. The other content is what he calls "cognitive content", which can be understood as bodies of information or dispositions to apply the concepts with the referential content it has.

[3] I take it that much of the current methodology in this area of ethics of AI implicitly follow such a method of focusing on application conditions. For example, Cappelen and Dever [6] engineer or propose

Footnote 3 (continued)

an externalist way of thinking about content that allows us to make sense of the output of machine learning algorithms without having to know much about their internal workings. Himmelreich and Köhler [15] mainly think about conceptual engineering of responsibly gaps in terms of assessing different intension and extension pairs in terms of how well they fulfill a certain function that we want the concept of responsibility to play.

not better or worse than any other intension or extension. Whether the extension of the word 'person' includes robots or not is by itself neither good nor bad. What gives conceptual engineering a *good rationale* is what follows from this in our community [20, 47]. However, determining the best *consequences* of application is usually the trickiest part. It often requires making difficult commitments and it is here where we can distinguish realistic from unrealistic proposals. For example, if we implement a concept of personhood that can include robots, does this mean that we are no longer able to sell certain robots given that persons cannot be bought and sold? Does this mean they now have the right to refuse to work with us, start a family and apply for a passport? For all the known robots, these consequences seem completely absurd. What could it mean for a robot to apply for a passport or start a family?

If we decided to change the consequences of applying the concept, this may even significantly change our concept of personhood as the concept of personhood might lose its ordinary function, arguably for the worse. If being a person no longer secures having rights (as would arguably be the case of robots), then this seems to seriously erode one of our most fundamental moral concepts. Thus, designing intensions and extensions is not enough. No practically relevant engineering proposal in the ethics of AI can do without a proposal regarding the consequences of the application of a word or concept.

4 A pragmatist model of conceptual engineering

4.1 An alternative to the representationalist approach

What does an alternative to the representationalist accounts look like? In the more recent literature (cf., [14], several pragmatist alternatives to representationalist accounts of the target objects of conceptual engineering have been developed that focus on the non-representational function of concepts [44–46], inferences and consequences of application [20, 43] as well as interpersonal entitlements and duties in concrete joint actions [24]. Generally speaking, these accounts have in common that they are skeptical of the role of language as merely serving a representational function, i.e., of merely representing reality. Instead, on a general level, pragmatist accounts of language take words and sentences to be tools or technologies that we use to interact with our social and natural environment (often explicitly inspired by American pragmatists like [3] or [40]. The role of conceptual engineers is then to design tools such that they help us *act* rather than mere representational devices. The role of the conceptual ethicist is to assess these tools regarding whether the conceptual activist should try to implement them.

But what can it mean for engineers in the ethics of AI to design tools that help us act better in accordance with our goals and values rather than represent the world? If conceptual engineers in the ethics of AI neither design robots nor representational devices, what do they design, such that it changes how we can act in the world? How can changing our language and concepts make us act differently? I take it that we can summarize pragmatist accounts of conceptual engineering such that they have as their target objects, not representational objects but interpersonal normative relations. Those relations regulate what kinds of things we are entitled to apply our words to and what follows from such applications, i.e., what it commits us to (cf., [3, 40]). In other words, according to some pragmatists of conceptual engineering and ethics, conceptual engineers have as their target objects entitlements and commitments to make certain applications and draw certain inferences concerning words and sentences. This approach not only considers the application conditions of a term but especially the consequences of application, which was something that opponents of representationalist approaches to conceptual engineering argue is missing from the received representationalist views.

To make things more concrete, what exactly does the intellectual exercise of designing entitlements and commitments to make applications and draw inferences have to do with acting in the world? The answer is that our individual actions are at all times heavily constrained not just by our physical abilities but by social norms. I am physically capable of crossing the street when the red light is on, but others will likely criticize me, and the police may even fine me. In other words, I am at all times constrained by norms. If I fail to comply with such norms, I risk becoming physically constrained (admitted to an institution against my will) or ostracized. Language constitutes at least a significant part of this system of norms that constrains our actions. Again, a popular way to illustrate this constraining but also enabling aspect of language is the concept of marriage. In 2022, same-sex marriage became legal nationwide in Mexico. Before that, homosexual couples were not able to get married in all parts of Mexico. A change in the law can be understood as a conceptual change or instance of conceptual activism. Conceptually engineering the official concept of marriage now commits the authorities to apply the concept of marriage to homosexual couples under certain conditions. In other words, it makes homosexual marriage possible.

Note that a commitment to the idea of words being normatively related to one another and our actions does not commit us to any claim about the essential or main function of language (cf., [20]. Even strict representationalists will admit that words have not only representational potential but can accomplish things in the world. To give

another example, if I tell you that my cat is pregnant, this slightly changes certain interpersonal normative relationships between us. You are now for example entitled to infer that I have a pet or that this pet is expecting babies. It would be odd or even wrong if I told you that I have a cat but not a pet. You, on the other hand, are now entitled to, for example, ask what I will do with the babies and perhaps even whether you can have one of them. This change in normative relations between the speakers, when utterances are made, is possible because words come equipped with a rich network of inferential relations to other words or concepts that members of the linguistic community are normally entitled to rely on when interacting with others. None of this means that these inferential relations necessarily determine the content of such concepts. This does not make them any less real.

4.2 A pragmatist framework for doing conceptual engineering in the ethics of AI

What then is the practice of conceptual engineering in the ethics of AI concretely according to the pragmatist (see Fig. 1)? In the remainder of this section, I propose a framework for thinking about the different steps or phases of conceptual engineering in the ethics of AI. I do not claim that this is the only or even the best way of thinking about the practice of conceptual engineering in the ethics of AI, but I take it to offer a helpful framework to begin thinking about what such a method could reasonably look like. This framework builds on the above-mentioned distinction between the different roles that conceptual engineering is associated with. I argued that conceptual engineering is often associated with three distinct stages or roles that can, however, be embodied by the same person—the designer role or design stage ("conceptual engineer"), the assessment role or assessment stage ("conceptual ethicist") and finally the activist role or activism stage ("conceptual activist"). Those are merely helpful labels or distinctions and should not be considered as anything but tools to make the different phases and activities often associated with the label "conceptual engineering" more salient.

First, I take it that a responsible conceptual engineer in the ethics of AI will get their assignments from the "conceptual ethicist", i.e., at the assessment stage. But what does the conceptual ethicist assess if not the engineered concept? I argue that the conceptual ethicist in the ethics of AI, at this stage, assesses our conceptual schema or parts of this schema rather than individual concepts. This means she first identifies room for improvement or *disruptions* [25] of our conceptual schemata [26], which were generated by new AI applications like a new chatbot, social robot, or a new disruptive algorithm. For instance, she might identify a *conceptual gap* (e.g., a new artificial artifact we cannot identify), a *conceptual overlap* (two concepts seem to apply to the AI and we do not know which one we should select), or a *conceptual moral misalignment* (the AI-generated conceptual uses that are not in line with our moral values). For an example of a conceptual overlap consider the question of whether a sophisticated chatbot can be a person, i.e., whether the concept of person or object applies to it (or some other concept). The idea is that both concepts seem to fit more or less and that we might have to decide which of those we decide to use. This generates a conceptual gap – it requires the introduction of a new concept. For an example of a conceptual misalignment consider the case where a certain technology changes the use of the concept of friendship in ways some may identify as misaligned with our values and principles.

At the stage of conceptual assessments, conceptual ethicists must have the necessary knowledge of stakeholders' interests and values, their moral principles, and the necessary factual beliefs, e.g., about the current or intended function of the concept. For example, the question of whether a robot should count as a friend can be analyzed in utilitarian or Kantian terms. The former might argue that we should classify the robot as a friend only if this generates positive all things considered well-being. The Kantian might reason from a very different point of view and may ground such questions in certain abilities to reason and autonomy, independently of whether the question would eventually generate the best outcome [31]. The conceptual ethicist of AI should also be aware of the current conceptual schema and what the respective concepts in this schema are contributing, i.e., what their intended and actual function is in the system and how it is related to the other nearby concepts. These different pieces of information constitute constraints for the conceptual engineer whose aim is to generate a new conceptual framework that is in a kind of *reflective equilibrium* [10]. According to Brun [4], p. 239, a *reflective equilibrium* is characterized by the idea that "judgements and principles are justified if judgements, principles, and background theories are in equilibrium; and this state is reached through a process that starts from judgements and background theories, proposes systematic principles and then mutually adjusts judgements, principles and possibly also background theories."[4]

The second step in the conceptual design process is the engineering or design process. The conceptual engineer will now propose to introduce/eliminate a concept or use of a word or to change the use of an established word or concept. A successful design proposal will take the constraints and goals of the conceptual ethicists into account and designs new conceptual tools that promise to best fulfill their role in this conceptual system. These constraints can be ethical,

[4] Note (against [15], pg. 60) that finding such an equilibrium can generate room for much revision, that may even include revising our most basic principles and theories.

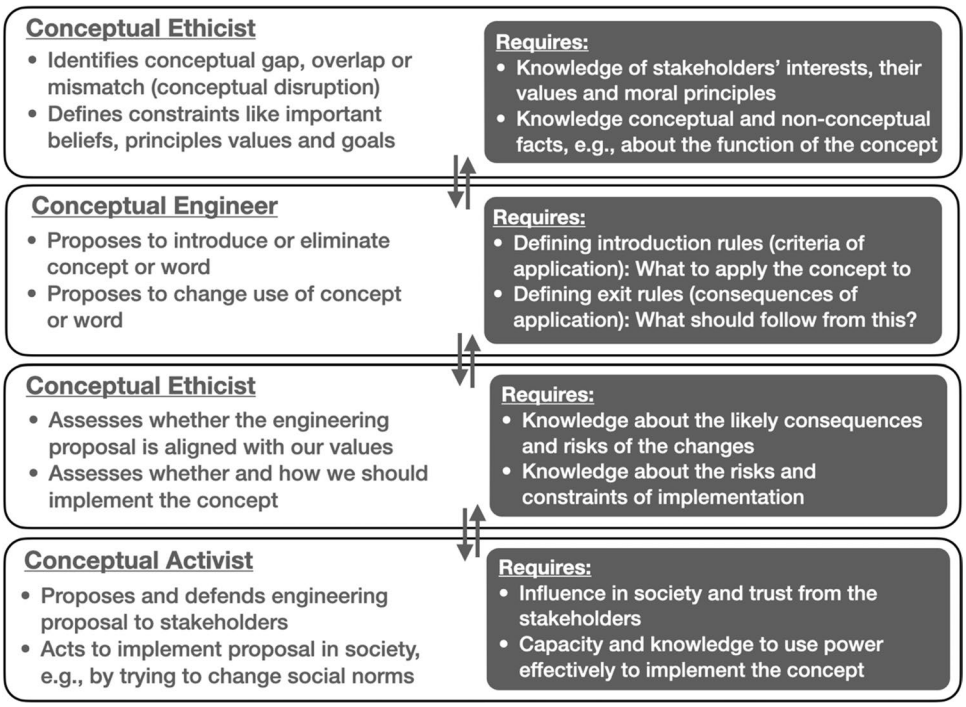

Fig. 1 Overview of different steps in a conceptual engineering project

epistemic, or prudential. For example, the ethicist may ask the engineer to design a concept that is as close as possible to the current use of a certain word or that captures nature at its joints. The aim is to make the job for the conceptual ethicist who gets again involved in the next stage of the design process as easy as possible. The conceptual engineer should propose or design therefore not only new conditions of application of an existing word or introduce such conditions for a new word, i.e., conditions of applications we can reasonably become committed to. She must then also determine what we are entitled or committed to infer from applications given that the proposed conditions of applications are satisfied. Such conceptual decisions are never risk-free and come with significant responsibility. The proposal must then also include a risk assessment that predicts possible risks regarding further conceptual disruptions [25] and unforeseeable inferential consequences. Ideally, several experts work out a different proposal based on the available empirical evidence, such that the conceptual ethicist can choose the best option available.

Now at the third stage of the design process, the conceptual ethicist gets involved again and will have to assess whether the proposed concept meets all the requirements (whether implementing the proposal would generate a reflective equilibrium). In addition, as [25] reminds us, the question of what concepts we choose (conceptual ethics) in an ideal world also has to be supplemented by another question on conceptual ethics. Assuming that one concept is better than another (one set of interpersonal entitlements to apply and conclude), when should we hand it over to the conceptual activist to lobby for implementing such changes? A concept can be better than others, but it might still be wrong to implement it if resistance to it is too large or the disruption caused by the intervention is too great [25]. A third important question is how we should implement an engineering proposal. Such an assessment of whether and how the proposal should be implemented requires again knowledge, not just about a specific function the concept should play but also about the likely consequences and risks of implementing the concept. It requires a lot of empirical knowledge, ideally modeling efforts, and ideally a team of researchers who can make such an assessment well.

Finally, if everything goes well, the ethicist hands it over to the conceptual activist who proposes and defends the engineering proposal to stakeholders. She will do so by trying to implement the proposal in society, e.g., by trying to change social norms. This can be done using public speeches (podcasts, interviews, TV appearances) or publications (books, articles, movies). This requires of course monetary means or influence in society and trust from the stakeholders. It ideally entails democratic legitimatization and avoids a form of conceptual paternalism [21, 35]. It also requires the capacity and knowledge to use power effectively to implement the concept. Again, each stage can be engaged in by the same team or even the same person. Such a team might include a publicist, sociologists, as well as philosophers. The idea is that the movement between the parts is dynamic, iterative, and non-linear (cf., [18]). Problems at the implementation stage may require a change in the overall engineering constraints and the proposal itself. For example,

if during the implementation process, it turns out that implementing the new concept of friend generates a lot of social disruption, the conceptual activist might report this back to the ethicist who asks the engineers to again change the design proposal.

5 Is conceptual engineering a novel method in the ethics of AI?

One might worry that the way I have been describing the practice of conceptual engineering in the ethics of AI, or at least what it should be like, is suspiciously close to what ethicists of AI are already doing. Once we concretely describe what exactly conceptual engineers are doing, and once we separate the practice of engineering from the practice of assessing the engineering, the resulting picture becomes very close to a re-description of what ethicists of AI are doing anyway—perhaps implicitly. First, finding a *reflective equilibrium* is clearly not a novel method in ethics (e.g., [36, 37]). Second, designing and assessing possible criteria and consequences of the application of words does not seem to be new either. At least in the ethics of AI, conceptual engineering seems to be exactly what ethicists of AI have been doing all along when they argue that the concept of, say, person, should apply to robots for the reason that such an attribution will have positive outcomes that are in line with our overall aims and principles. Opponents of such a proposal often respond by giving reasons against such an inclusion by referring to different or the same values that they take to be inconsistent with this proposal. This debate is at least aimed at getting closer to some kind of equilibrium that accounts for all our interests and commitments.

At this point, you might object that I have argued above that conceptual engineering proposals should include proposals as to the consequences of application and that this is not what many ethicists of AI currently are doing—they only propose new conditions of application—not what will follow from such applications. If this is correct (certainly it will depend on the respective author and paper), all this shows that these engineers are only doing half the job *explicitly*. The rest of the job is always, however, implied. Put differently, if they fail to engineer the consequences of application explicitly, they are bound to do so tacitly. Ethicists of AI then simply assume that the inferential consequences should remain the same. And, as argued above, this is often not possible if the conditions of application but not the consequences of application apply to the recently included kinds. To put this idea more succinctly: representational approaches to conceptual engineering, too, are engineering conditions that, if implemented, would entitle the speaker to apply a word to an object. However, they commit themselves always also to the normative consequences of application *remaining the same*. Ideally, however, such consequences should also be designed or *explicitly* stated as part of the engineering proposal.

For example, consider a representational approach to the question of whether a robot or chatbot can be a colleague (cf., [31], pp. 227). The currently most common approach to such conceptual questions in the ethics of AI is to first recover a set of application conditions of the respective word that has been proposed in general philosophical approaches. Nyholm for example relies on Betzler and Löschke's [2] set of application conditions (intensions) for applying the term colleague. Colleagues on their account are employed at a similar hierarchical level (they assume that a low-ranking soldier and a general are not obviously colleagues), engage in similar tasks or different tasks toward the same goal, and can judge one's success or failure better than one's friends and family. Nyholm then argues that these conditions do not apply to the robots we currently have and therefore concludes that robots of 2023 cannot be colleagues. If we construe this as a kind of closeted conceptual engineering process—trying to find application conditions that match the ones close to our ordinary practices—the assumption is that once we apply or don't apply the concept, the consequences of application are the same. Since Nyholm does not engineer them, he implicitly assumes that they should remain the same. But this assumption might not be shared if we believe that our current use of the word *colleague* might not be the ethically best use (which, again, Nyholm assumes but does not argue for).

But how then does conceptual engineering differ from a more realist approach of trying to describe real kinds or a conceptual analysis approach? In other words, an immediate objection to my claim could be that most ethicists of AI currently view themselves as engaging in a form of realist project or as engaging in conceptual analysis. Conceptual engineering is usually introduced as the practice of designing the concepts we *should* use rather than the concepts we are *currently* using. I take it that these other practices can be understood as engineering constraints grounded in additional constraints given by the conceptual ethicist. Conceptual analysis, too, can simply be construed as a constraint on our engineering projects. This constraint of a conceptual analysis project is trying to implement a concept that is as close as possible to our ordinary concept. A realist project can be understood as introducing the constraint that the engineering product should capture the joints in nature. But the constraints are part of conceptual ethics and not conceptual engineering. The conceptual engineer simply argues for conditions and consequences of application that meet the constraints set by the conceptual ethicist.

So, is it all hype? Is conceptual engineering not worth the attention it receives? Assuming that the literature on

conceptual engineering has only recently found its way into the ethics of technology and AI, we might also ask whether we should give it any attention at all in this community. Or should we simply sit this one out? I do not think that the result that conceptual engineering might be merely what we have been doing all along is best understood as a negative result and that reflecting on our methods in the ethics of AI is incredibly important, especially now that AI is disrupting many of our conceptual practices. I, therefore, take it to be in line with a recent trend in meta-metaphysics that tries to make sense of disagreements in metaphysics. Assuming that ethics of AI has a metaphysical component (what are questions like what it is to be a robot, can robots be persons or friends), this then merely results in the idea that describing the method in ethics of AI as conceptual engineering can make more sense of this field. In this picture [42], metaphysics in the ethics of AI can be understood as debates on how to use language. One charge against such a view is that such a linguistic interpretation of the practice cannot make sense of the fact that we take this practice often to be rather deep and consequential as opposed to merely a mere linguistic disagreement, i.e., a disagreement about words.

However, such linguistic disagreements are as we have seen above far away from inconsequential. Whether we engineer a concept of person or friend that comes as close as possible to the current concept or whether we introduce a concept of friendship or person that includes robots will have real-world consequences. If we decide that a robot can be a friend and we keep what follows from such an application constant, then this means that we are entitled to expect loyalty from the robot and that the robot can expect loyalty from us. We cannot call ourselves a friend of something that we would immediately sell as soon as the new model comes out. But this sounds absurd. Given the absurdity (the difficulty to even make sense of this possibility), would this then mean that we have to change our classification behavior of objects and person as well—change those inferentially related concepts as well? This all will depend heavily on our values and goals and people with different values and goals likely get to different conclusions, which will likely explain their apparent deep metaphysical disagreements.

A second reason why I believe that the finding that conceptual engineering is what we have been doing all along is not a negative result is that it is an opportunity to make our methodological commitments explicit and even to improve our methodology. It will enable us to better identify conceptual issues that we might have taken for granted and to identify reasons why they so easily generate debates and disagreements. It may even lead us to generate a general model of how to best do ethics of AI at least when it comes to conceptual issues like whether a robot can be a person or our friend. Above, I have developed the beginnings of such a model and I hope that at least some of it might be implemented in future projects in the ethics of AI. A recent paper by Veluwenkamp and Van den Hoven [46], too, developed a different but compatible model of how to do ethics of AI properly. They apply the more general approach of Value Sensitive Design to the conceptual engineering process. In other words, even if conceptual engineering is something we have been doing all along in the ethics of AI, now, we have the opportunity to do things properly, i.e., in an empirically informed and systematic goal-oriented, and value-sensitive way. We have the opportunity to engineer our concepts more rigorously and systematically.

Conceptual engineering in the ethics of AI then probably requires a team of researchers of different skill sets. Value-sensitive design of technologies is difficult. It requires the identification of direct and indirect stakeholders and their values, which is at least to a significant degree an empirical project, which can only be done properly by including sociologists or by introducing empirical methods (experimental philosophy) of the ethics of AI. We then need to sharpen these value terms into operationalizations or more specific concepts, which, I assume, is primarily a job for philosophers. What for example we mean by the value-term *privacy* requires reflection and knowledge of the literature on privacy. How do we then turn these values into concrete constraints for conceptual design? I take it that this is the responsibility of a conceptual ethicist again and little that an empirically working researcher and contribute much to. Finally, however, we need to test our conceptual designs and see if they really are taken up by the community and if they are really in line with our principles and values and the constraints of the conceptual ethicist. This is again an empirical question, as well as a project for conceptual activists. Thus, I conclude that the ethics of AI program seems to be a much more ambitious one, requiring an interdisciplinary approach and team, if we want to do it properly.

6 Conclusion

The upshot of this paper is simple. It is not clear what conceptual engineering in the ethics of AI is supposed to be. The currently most popular approaches in the core conceptual engineering debate are representationalist accounts, which do not seem to be what we are or ought to be doing as ethicists of AI. Instead, I argued that the conceptual engineer in the ethics of AI is best understood as proposing inferential relations (application conditions and consequences of application) that we could become entitled to. The ethicist then evaluates these choices and assesses whether they ought to be implemented. If the ethicist gives their permission, the project is handed over to the activist who aims to implement the inferential relations for example by trying to change the relevant social norms. How does the assessment

of inferential relations and questions of implementation work? I argued that this works most likely via aiming at deriving a certain reflective equilibrium. The conditions and consequences of the application of a word must be in the proper relationship with our overall beliefs, principles, and goals. This method of engineering inferential networks and assessing them however does not constitute a novel method. The method of finding a reflective equilibrium is not new. Conceptual engineering in the ethics of AI is then simply the meta-metaphysical acknowledgment of a philosophical method rather than the invention of a new method.

Data availability The analysis presented in this review article is purely theoretical in nature and does not involve the use of specific datasets.

Declarations

Conflict of interest I have no competing interests to report.

References

1. Babushkina, D.: What does it mean for a robot to be respectful? Techné Res. Philos. Technol. **26**(1), 1–30 (2022)
2. Betzler, M., Löschke, J.: Collegial relationships. Ethical Theory Moral Pract. **24**, 213–229 (2021)
3. Brandom, R.: Making it Explicit: Reasoning, Representing, and Discursive Commitment. Harvard University Press, Cambridge (1994)
4. Brun, G.: Explication as a method of conceptual re-engineering. Erkenntnis **81**(6), 1211–1241 (2016)
5. Cappelen, H.: Fixing Language: An Essay on Conceptual Engineering. Oxford University Press, Oxford (2018)
6. Cappelen, H., Dever, J.: Making AI Intelligible: Philosophical Foundations, p. 192. Oxford University Press, Oxford (2021)
7. Dobler, T.: Pluralist conceptual engineering. Inquiry (2022). https://doi.org/10.1080/0020174X.2022.2086171
8. Floridi, L.: The Fourth Revolution: How the Infosphere is Reshaping Human Reality. OUP, Oxford (2014)
9. Floridi, L.: The Logic of Information: A Theory of Philosophy as Conceptual Design. Oxford University Press, Oxford (2019)
10. Goodman, N.: Fact, Fiction, And Forecast, 4th edn. Harvard University Press, Cambridge (1983)
11. Grice, H.P.: Meaning. Philos. Rev. **66**, 377–388 (1957)
12. Gunkel, D.J.: How to Survive a Robot Invasion: Rights, Responsibility, and AI. Routledge, New York (2019)
13. Haslanger, S.: Resisting Reality. Oxford University Press, Oxford (2012)
14. Henne, C., Huetter-Almerigi, Y.: Conceptual engineering and pragmatism: historical and theoretical perspectives. Inquiry (2022). https://doi.org/10.1080/0020174X.2022.2158927
15. Himmelreich, J., Köhler, S.: Responsible AI through conceptual engineering. Philos. Technol. **35**(3), 1–30 (2022)
16. Isaac, M.G.: How to conceptually engineer conceptual engineering? Inquiry (2020). https://doi.org/10.1080/0020174X.2020.1719881
17. Isaac, M.G.: What should conceptual engineering be all about? Philosophia **49**(5), 2041–2051 (2021). https://doi.org/10.1007/s11406-021-00367-x
18. Isaac, M.G., Koch, S., Nefdt, R.: Conceptual engineering: a road map to practice. Philos. Compass **17**(10), e12879 (2022)
19. Jackson, F.: From Metaphysics to Ethics: A Defence of Conceptual Analysis. Clarendon Press, Oxford (1998)
20. Jorem, S., Löhr, G.: Inferentialist conceptual engineering. Inquiry (2022). https://doi.org/10.1080/0020174X.2022.2062045
21. Kitsik, E.: Epistemic paternalism via conceptual engineering. J. Am. Philos. Assoc. **1**, 20 (2022). https://doi.org/10.1017/apa.2022.22
22. Koch, S.: Engineering what? On concepts in conceptual engineering. Synthese **199**(1–2), 1955–1975 (2021). https://doi.org/10.1007/s11229-020-02868-w
23. Koch, S., Löhr, G., Pinder, M.: Recent work in the theory of conceptual engineering. Analysis, 1–15 (forthcoming)
24. Löhr, G.: Commitment engineering: conceptual engineering without representations. Synthese **199**(5), 13035–13052 (2021)
25. Löhr, G.: Linguistic Interventions and the Ethics of Conceptual Disruption. Ethic Theory Moral Prac **25**, 835–849 (2022). https://doi.org/10.1007/s10677-022-10321-9
26. Löhr, G.: Do socially disruptive technologies really change our concepts or just our conceptions? Technol. Soc. (2023). https://doi.org/10.1016/j.techsoc.2022.102160
27. Machery, E.: Doing without concepts. Oxford University Press (2009)
28. Machery, E.: Philosophy Within Its Proper Bounds. Oxford University Press, Oxford (2017)
29. Manne, K.: Down girl: The logic of misogyny. Oxford University Press (2017)
30. McPherson, T., Plunkett, D.: Conceptual ethics and the methodology of normative inquiry. In: Conceptual Engineering and Conceptual Ethics, pp. 274–303 (2020)
31. Nyholm, S.: This is Technology Ethics. Wiley-Blackwell, Hoboken (2023)
32. Nyholm, S., Smids, J.: Can a robot be a good colleague? Sci. Eng. Ethics **26**(4), 2169–2188 (2020)
33. Pinder, M.: Conceptual engineering, metasemantic externalism and speaker-meaning. Mind **130**(517), 141–163 (2021)
34. Pollock, J.: Content internalism and conceptual engineering. Synthese **198**(12), 11587–11605 (2021)
35. Queloz, M., Bieber, F.: Conceptual engineering and the politics of implementation. Pac. Philos. Q. **103**(3), 670–691 (2022)
36. Rawls, J.. The independence of moral theory. In: Collected Papers, pp. 286–302. Harvard University Press, Cambridge (1999)
37. Raz, J.: The claims of reflective equilibrium. Inquiry **25**(3), 307–330 (1982)
38. Sawyer, S.: Concept pluralism in conceptual engineering. Inquiry (2021). https://doi.org/10.1080/0020174X.2021.1986424
39. Scharp, K.: Replacing Truth. OUP, Oxford (2013)
40. Sellars, W.: In the Space of Reasons: Selected Essays of Wilfrid Sellars. Harvard University Press, Cambridge (2007)
41. Simion, M.: The 'should' in conceptual engineering. Inquiry **61**(8), 914–928 (2018)

42. Thomasson, A.L.: Ontology Made Easy. Oxford University Press (2014)
43. Thomasson, A.: Conceptual engineering: when do we need it? How can we do it? Inquiry (2021). https://doi.org/10.1080/0020174X.2021.2000118
44. Thomasson, A.L.: A pragmatic method for conceptual ethics. In: Conceptual Engineering and Conceptual Ethics, pp. 435–458 (2020)
45. Veluwenkamp, H., Capasso, M., Maas, J., Marin, L.: Technology as driver for morally motivated conceptual engineering. Philos. Technol. **35**(3), 1–25 (2022)
46. Veluwenkamp, H., van den Hoven, J.: Design for values and conceptual engineering. Ethics Inf. Technol. **25**(1), 1–12 (2023)
47. Weinberg, J.: What's epistemology for? The case for neopragmatismin normative metaepistemology. In: Hetherington, S. (ed.) Epistemological Futures, pp. 26–47. Clarendon Press, Oxford (2006)