Responsible AI through Conceptual Engineering

Johannes Himmelreich, Syracuse University

Sebastian Köhler, Frankfurt School of Finance & Management

Abstract

The advent of intelligent artificial systems has sparked a dispute about the question of who is responsible when such a system causes a harmful outcome. This paper champions the idea that this dispute should be approached as a *conceptual engineering problem*. Towards this claim, the paper first argues that the dispute about the responsibility gap problem is in part a *conceptual dispute* about the content of RESPONSIBILITY and related concepts. The paper then argues that the way forward is to evaluate the conceptual choices we have, in the light of a systematic understanding of why the concept is important in the first place—in short, the way forward is to engage in conceptual engineering. The paper then illustrates what approaching the responsibility gap problem as a conceptual engineering problem looks like. It outlines argumentative pathways out of the responsibility gap problem and relates these to existing contributions to the dispute.

0. Introduction

Artificial Intelligence (AI) is progressing. Autonomous weapons systems (AWS) and autonomous vehicles (AV) are getting ever closer to entering widespread use. This prospect has sparked a dispute surrounding the question of who is responsible when the use of such technologies causes harmful outcomes. Whether it is the use of AVs, AWS, or other AI applications, there are some outcomes for which it seems that *no one* would be responsible. This is the *responsibility gap problem*. To make matters worse, this problem is bound to occur in a wide range of cases (see e.g., Coeckelbergh, 2016, Danaher, 2016, Gunkel, 2017, Hevelke & Nida-Rümelin, 2015, Hellström, 2013, Himmelreich, 2019, Köhler, 2020, Köhler et al., 2017, Liu, 2017, Matthias, 2004, Nyholm, 2018, Robillard, 2018, Roff, 2013, Sparrow, 2007, or Tigard, 2020).

We argue that the responsibility gap problem should be approached as a *conceptual engineering problem*. To make progress on the question of whether there is a responsibility gap, conceptual questions of RESPONSIBILITY and related concepts must be investigated systematically.[1] Specifically, such an investigation should be reformative. It should reflect on the existing conceptual repertoire and ask how it could be *improved*. It asks not only: What roles does

[1] We follow the convention of using small caps to refer to concepts, to distinguish concepts from their extension as well as from the expressions referring to them.

This is a draft. The version of record, published in *Philosophy & Technology* can be found here: https://doi.org/10.1007/s13347-022-00542-2

RESPONSIBILITY play in reasoning? But also: What role *should* it play, what could and should be its content? That is, what *conception* of RESPONSIBILITY ought to be used? A conceptual engineering approach assumes that there are many possibilities on what the content of a given concept could be: There are many possible ways of making the content of a concept precise without a change in topic.[2] Conceptual engineering also assumes that what the precise content of a concept ends up being is, in some way, up to us, the concept users. What precise content we give to a concept depends on answers we give to certain questions we must ask to make the concept's content more precise. For example, does RESPONSIBILITY presuppose control? Answering such questions in one way or another yields a different content for RESPONSIBILITY. Giving such an answer is what we call a "conceptual choice."

That there *are* such conceptual choices that determine the content of our concepts and that we need to think about the choices we *ought to take* is a basic assumption of conceptual engineering; though authors differ on what, exactly, this amounts to, since semantic internalists see the nature of conceptual choices very differently from semantic externalists. We do not take a stance on the nature or mechanisms of conceptual choices. Our argument does not depend on it and we thus remain neutral on substantial questions where we can (see e.g. Cappelen, 2018; Thomasson, 2021). In fact, our aim in this paper is *not* to defend the basic assumptions of conceptual engineering. What concepts are, what conceptual engineering is, whether it is possible, and how it could and should be done—these are questions that are widely discussed elsewhere in the literature, especially in the last few years (see e.g. the discussions in Cappelen, 2018; Cappelen et al., 2020). In this paper we, instead—given the basic assumptions of conceptual engineering—argue that conceptual questions are central to the responsibility gap dispute and we illustrate how conceptual engineering could proceed in this dispute.

Of course, the dispute on the responsibility gap problem has already brought conceptual questions into focus. For example: Some assume that control is necessary for RESPONSIBILITY, and investigate the content of CONTROL (e.g. de Sio & van den Hoven, 2018, Himmelreich, 2019). Others home in on the concept of AGENCY and debate whether AWS and AVs have it and how it relates to human agency (e.g. Nyholm, 2018; Robillard, 2018). Again, others have argued that novel concepts must be introduced (e.g. Hellström, 2013, Pagallo,

[2] We assume that the same concept can have different precise contents (as we explain later). Not everyone accepts this assumption. Instead, on a common view concepts are individuated by their content. However, our argument is compatible with this view: Our assumption can be rephrased in terms of "topics" instead of "concepts"—and what we call "conceptions" would then be called "concepts" (see Cappelen, 2018: 107-121). The exact formulation of these assumption will not matter for what we argue here. Everything we say should, *mutatis mutandis*, be applicable no matter the exact details of the conceptual engineering approach one chooses.

2011). But even when conceptual questions have come into focus, they are rarely conceived as involving an *evaluation* of the different possibilities of conceptual content. In result, the existing dispute lacks in methodological reflection and intention. Our claim—that the responsibility gap *is* and *should be approached* as a conceptual engineering problem—thus aims to reorient the dispute towards addressing conceptual questions as conceptual choices. Such a conceptual engineering approach is particularly sensible when social or technological advances pose problems for which our current conceptual repertoire might no longer be ideally suited—such as the responsibility gap problem.

Seeing the responsibility gap as a conceptual engineering problem advances the dispute in at least two significant ways. First, it transcends first-order disputes. We argue that the question of whether there *are* responsibility gaps should be set aside. Any answer to this question depends in large part on the concepts involved, such as RESPONSIBILITY or CONTROL. Second, it shifts the focus onto a principled approach to evaluating different possibilities of conceptual content This brings important philosophical-methodological questions into the foreground: Why are these concepts important in the first place, what do they do for us that is important—what is their function? And, it highlights that we should approach our concepts systematically from that perspective, considering how our concepts can do best what is most important.

What we propose does not amount to a solution to the responsibility gap problem. Neither is this our aim. Rather, our aim is programmatic. We argue that the contributions to the responsibility gap dispute should engage explicitly with normative questions concerning conceptual choices. We point out these choices and develop and illustrate a framework of how to approach them. In this sense, this is a paper on the meta-philosophy of the responsibility gap problem. Whether or not there is a responsibility gap, is not a question of this paper. We follow the literature on the responsibility gap problem in assuming that there *is* a responsibility gap problem and that this problem involves the concept of RESPONSIBILITY. Whether the responsibility gap problem could be stated without speaking of "responsibility" or without employing the concept of RESPONSIBILITY, is not a question of this paper.

We pursue our programmatic aim pragmatically: We provide the starting resources needed to approach the responsibility gap as a conceptual engineering problem. Specifically, we illustrate what new directions the debate about the responsibility gap could take, what kinds of issues the conceptual engineering approach highlights, and what issues have to be investigated. We review some existing contributions and what stance they take on these conceptual issues. We also identify different functions that RESPONSIBILITY can play—such as a "desert function," a "ledger function" or an "incentive function." The conceptual engineering approach to the responsibility gap then consists in the question: Can RESPONSIBILITY perform its most important functions just as well (or better) if it is engineered to avoid responsibility gaps? We illustrate what answering this question looks like by comparing a view that offers conceptual choices that close responsibility gaps with one that generates such gaps in the light of the functions of RESPONSIBILITY we identify.

The paper proceeds as follows: A first order of business is to get a clear understanding of the first-order dispute. Section 1 presents the responsibility gap problem and what is at stake in this debate. Section 2 then argues, by way of cursorily surveying the literature, that the dispute about the responsibility gap problem is partly a *conceptual dispute*. That is, disagreements about who is responsible are grounded, at least to a significant extent, in disagreements over the content of underlying concepts such as RESPONSIBILITY or AGENCY. Section 3 then argues that the way out of this conceptual dispute is to approach the responsibility gap problem as a conceptual engineering problem. Section 4 illustrates what approaching the responsibility gap problem as a conceptual engineering problem looks like, by identifying a list of functions RESPONSIBILITY plays and assessing the conceptual choices made by some views in the literature in the light of these functions.

1. What is the Responsibility Gap Problem?

With the success of machine learning (ML) techniques and the wealth of available data, it is becoming possible to build increasingly sophisticated systems, which perform ever more complex tasks. The responsibility gap problem arises when these systems become so advanced that they can, plausibly, be said to make decisions—that they, in this sense, become *agents*.[3] These are systems capable of gathering and processing information, assessing such information in the light of the goals set for them, and making and executing decisions based on such assessment. After a relevant training phase, such systems can be expected to perform a range of tasks not only expertly but *autonomously* in the sense that they can execute them without human interference,[4] while exhibiting purposeful or complex decision-making that can adapt and function in some—albeit limited—range of circumstances. AWS and AVs are clear examples of such systems. Other examples are ML classifiers that determine whether a claim for unemployment

[3] See e.g. Köhler, 2020: 3124/3125; Nyholm, 2018, for attempts to flesh out what it might mean to call these systems "agents"; note that this is also a subject for conceptual engineering.

[4] So, they are only "autonomous" in the sense used in robotics [e.g. Beer et al., 2014; US Department of Defense, 2012], not in a more robust philosophical sense (e.g. Hooker & Kim, 2019; Totschnig, 2020).

insurance is eligible or whether a picture contains the face identical to that of a known terrorist, medical systems that diagnose cancer, or health care robots—each time an AI either acts itself or significantly contributes to the practical reasoning of another agent. For the purposes of this paper, let us call such systems "artificial intelligence" (AI).

For AI, the responsibility gap problem arises as follows: Suppose something *does* go wrong when an AI decides. Specifically, assume that the system makes a decision that causes some harm.[5] Assume furthermore that neither the kind of failure that resulted in the harm, nor the harm itself, could have been foreseen by anyone. The AI system had been carefully developed and diligently tested. Assume also that it is not a harm that was intended by those who designed or used the system (for example, some of the harms that AWS's decisions cause are intended by those who deploy them). This kind of case raises a crucial question: Who is morally responsible for this harm? This question leads straight into the responsibility gap problem (e.g. Danaher, 2016; Matthias, 2004; Roff, 2013; Sparrow, 2007).

The problem arises for two reasons. First, it seems that the AI itself cannot be morally responsible for the harm, because (at least in the foreseeable future) no AI is, plausibly, a *moral* agent (see e.g. Fossa, 2018; Hakli & Mäkelä, 2019; Hew, 2014; Himma, 2009; Véliz, 2021). Second, it appears that all humans involved in the situation fail at least one necessary condition for attributing moral responsibility due to the distinct agency of the AI.[6]

Human responsibility may be undermined in different ways. For example, human responsibility could be undermined because of the *epistemic condition*—that is, because of what the human could reasonably foresee—or because of the *intentional condition*—that is, because of what the human intended or something about their "quality of will". We concentrate on the *control condition* (Fischer & Ravizza, 1998: 12). That is, we concentrate on how AI's agency may undermine the human's control of the right kind and, thereby, their responsibility.[7]

Consider the following statement by Danaher (2016: 301):

[5] We use the term "harm" here broadly, to not just cover individual, but also collective harms. For example, if a hiring algorithm decides to not hire someone because of their race, we would count such a case as relevant, even if that person is then hired for an even better job elsewhere and so is not, in the ordinary sense, harmed by that decision. We also only consider harms, for the purposes of this paper, that are unjustified.

[6] Responsibility gaps may arise not only for AI agents but also for group agents. Yet, the literature on group agency largely rejects the claim that there are significant collective responsibility gaps (Braham & van Hees, 2010: ch. 7; Collins, 2019; Duijf, 2018; List & Pettit, 2011) (see e.g. List & Pettit, 2011: ch. 7).

[7] This is where the responsibility gap for AI differs from that of group agency. Whereas in the case of AI, the AI's agency undermines human control, in the case of group agency, human agency undermines the group's agency.

> A robotic agent, with the right degree of autonomous power, will tend to be causally responsible for certain injurious or harmful actions. However, the robot will not be morally and legally responsible (because it will lack the requisite moral capacities), nor will the human creators and designers be morally/legally responsible because the robot has a sufficient level of independence from them.

The argument proceeds as follows. Moral responsibility presupposes sufficient control: One can be responsible for an outcome only if one has sufficient control over that outcome. Danaher likely invokes this idea with the expressions of having "autonomous power", being "causally responsible", and having "a sufficient level of independence". A sufficiently advanced AI will have control in this sense. But, in turn, no *human* has sufficient control over harmful outcomes in the relevant sorts of cases, "because the robot has a sufficient level of independence". In short, the independent *agency* of the AI interferes with human control. Therefore, no human could be morally responsible for such outcomes. Presumably, though, the AI itself cannot be responsible either, as it lacks certain relevant capacities. Yet, if neither the AI nor any human are responsible for the relevant harmful outcomes, it seems that *no one* would be responsible for them: there would be a responsibility *gap*. This is the responsibility gap problem. Almost identical arguments have been made by others in the literature (e.g. Matthias, 2004; Roff, 2013; Sparrow, 2007).

Much of the dispute around the responsibility gap problem has concentrated on AWS and AV (e.g. Burri, 2017; Danaher, 2016; Hevelke & Nida-Rümelin, 2015; Matthias, 2004; Nyholm, 2018; Roff, 2013; Schulzke, 2013; Sparrow, 2007; Vladeck, 2014). But nothing singles out AWS or AV as particularly relevant. Rather, the responsibility gap problem arises for *any AI* that engages in decision-making regarding tasks that could, potentially, have harmful outcomes. Arguably *most* tasks that will be taken over by AI fall into this category. So, the responsibility gap is a general problem for AI, not just one that arises for AWS or AV.

What is Wrong With Responsibility Gaps?

The relevance of the debate on AI responsibility gaps hangs on the question of why, if at all, responsibility gaps are morally problematic. After all, there are many harmful outcomes for which nobody is responsible. An erupting volcano, an earthquake, or a meteor strike each are harmful and (generally) nobody is responsible for them, but they are not morally problematic in the same way that responsibility gaps are taken to be.

We see three main reasons why responsibility gaps are morally problematic. First, responsibility gaps conflict with near-universal pre-theoretic moral judgments or sensibilities.

Most people feel that someone *should* be held responsible for the harm caused by AI. After all, the situations that result in the harm do *not* resemble paradigmatic examples of pure accidents or acts of nature. Quite the opposite: Given the ways in which humans are involved in the design, testing, building, and deployment of AI, and given the fact that there are people who *benefit* from the employment of the AI, the situations involving AI strongly resemble those in which humans use artifacts. This strong resemblance to situations of *human-made* harm supports the judgment that someone is responsible. Insofar as such pre-theoretic moral judgments should epistemically guide moral assessment, this suggests that responsibility gaps are *morally problematic*. Responsibility gaps create a "normative mismatch" (Köhler et al., 2017: 54).

A clear example of such a mismatch are AWS. Responsibility gaps for AWS might undermine justice in war, in line with Michael Walzer's dictum that "there can be no justice in war if there are not, ultimately, responsible men and women" (Walzer, 1977: 287; see also Sparrow 2007: 67). Moreover, the rules of *just war* may require holding those who harm non-combatants responsible. Otherwise, if no one can be responsible for the harm created by AWS, the increasing use of AWS would create and perpetrate injustice.

Second, responsibility gaps may undermine accountability in public institutions. AI will find increasing use in administrative decision-making, both for gathering and processing information, but also for automating certain kinds of decisions (Bullock, 2019). When decisions are made in public institutions, e.g. in government, civil service or local administration, it is an important *democratic* requirement that politicians, public administrators, and civil servants can be held accountable for these decisions and their outcomes (Lechterman, 2022). Unfortunately, the use of AI may lead to responsibility gaps. Insofar as accountability implies responsibility, responsibility gaps are accountability gaps.

Third, responsibility gaps might limit the uptake of AI and thereby make it harder to obtain large potential increases in social welfare. For one, based on the first two points on why responsibility gaps are morally problematic, some call for a ban of certain uses of AI (e.g. Sparrow 2007; Campaign to Stop Killer Robots (2021)). Moreover, responsibility gaps can lead to a wide-spread *mistrust* of technology. This mistrust could potentially dampen innovation and may lead to a slower proliferation of AI. This is a problem insofar as AI may be hugely beneficial. AI might make traffic safer, information gathering and processing capabilities more powerful, administration more efficient (for an example from criminal justice, see Kleinberg et al., 2018). Likewise, AWS could be on average *less* harmful than conventional weapons (e.g. Burri 2017, Müller & Simpson 2016). A ban or a significantly slower uptake of AI may deprive societies of such benefits and, therefore, come with its own significant moral cost.

Considering these stakes, we turn to the crucial question: Are there AI responsibility gaps?

2. AI & Responsibility: A Conceptual Dispute

Whether there are responsibility gaps for AI depends, in crucial parts, on how central concepts— RESPONSIBILITY, AGENCY, or CONTROL—are understood. The responsibility gap dispute is, hence, what we call a "conceptual dispute". That is, it is a dispute over a question that, in order to be fully answered, requires—among many other things—a view about the precise content of one or more concepts that are relevant to the dispute. The claim that the responsibility gap dispute is such a conceptual dispute is the claim that we defend in this section. We survey the literature and find that the question of whether there are AI responsibility gaps has been answered in the negative as well as the affirmative, but that each answer depends, crucially, on choices about the content of concepts involved. We present one partial diagram that represents some of the conceptual choice points we observe in the literature (Figure 1). First, though, we clarify what a conceptual dispute is by way of an example.

Conceptual Disputes

Without assuming any particular view or theory on what concepts are, we take it as a basic assumption that many concepts are imprecise or indeterminate.[8] Such concepts allow for several conceptual possibilities, that is, there are different ways of filling out their content and making them precise.

None of this is controversial in any way. This picture is consistent with a mainstream approach to philosophy, the analytical tradition, which contends that (philosophical) questions can be answered only after prior questions about the relevant concepts and their content have been answered. In debates on free will in analytic philosophy, we, hence, encounter questions such as: Does FREE WILL require an ability to do otherwise? Does FREE WILL require some kind

[8] This might not be true of all concepts, but, as we argue below, it is consistent with the dispute about RESPONSIBILITY, on which we focus.

of control—that is, is the concept of FREE WILL such that its correct application requires that a certain other concept, CONTROL, applies?[9]

We call such questions—each of which concerns a way of making indeterminate content of a concept precise—*conceptual questions.*[10] The examples above are conceptual questions about FREE WILL: They are questions about what, exactly, could be meant by "free will" when the question is whether free will is compatible with determinism. That this question leads to questions about CONTROL shows that answering questions about FREE WILL, of course, raises further conceptual questions about other concepts.

There is a distinction between a concept's content and its extension. Philosophy is concerned not only with the former but also the latter, with concepts' extensions. A question of content is what FREE WILL requires. A question of extension is whether free will actually exists, that is, whether anything in the actual world falls into the extension of FREE WILL. In philosophical disputes, questions of content and extension go hand in hand. To answer whether free will actually exists—whether "free will" refers to anything in the actual world—it is necessary to precisify the content—what is FREE WILL. For our purposes here, we mean by "conceptual questions" only questions about content but not about extension.

[9] A reviewer asks: Why think that these are questions about the *concept* FREE WILL as opposed to the *property* of free will? We do not rule out that conceptual questions about FREE WILL are questions about the property. But this paper sets questions of ontology aside. We do not assume any theory, let alone any ontology, of concepts. Properties, however, are an ontological category. However, properties are sufficiently similar to concepts to not rule out interpreting concepts as properties—for example, properties, like concepts, have an extension: the set of its instantiations. In fact, on one view of concepts—seeing concepts as abstract objects—concepts are properties. Thus, on this view, questions about the concept of FREE WILL would be questions about the property of free will.

[10] A reviewer asks: Why think that these are *conceptual* questions? This question raises significant meta-philosophical issues that reach far beyond the scope of this paper. We can give two brief motivating arguments. First, these questions are conceptual questions because they are questions about constituents of our thoughts (i.e. concepts). These are the kinds of questions we ask ourselves when we unpack what we mean by "free will", which names a constituent of our thought. Second, we take it as a basic fact that, by definition, questions of the form "what is *x*?" are conceptual questions. The question "does FREE WILL require and ability to do otherwise" might not have the same surface grammar "what is free will?", but the former regularly arises in attempts to answer the latter. Questions of this latter form probe what something must be like to fall under the concept, which is exactly what the former question ("what is *x*?") is after.

Conceptual questions have answers. Let us call a determinate set of conceptual choices regarding the content of a concept a *conception* of the concept.[11] A conception of a concept is, as we understand it here, a possible way of making the content of the concept precise.

A *conceptual dispute* is a dispute over a question that is grounded, at least to a significant extent, in disagreements over the content of one or more underlying concepts.[12] Different participants to a dispute might operate with different conceptions of a concept. This definition of a conceptual dispute draws on Chalmers' (2011) work on verbal disputes. On the one hand, Chalmers (2011) sees verbal disputes as marred by a "familiar and distinctive sort of pointlessness" (525), on the other, he also says that "the diagnosis of verbal disputes [is] a tool for philosophical progress" (517).

We, as proponents of conceptual engineering, see conceptual disputes in this latter spirit. A conceptual dispute is an opportunity to move philosophical discussions forward. This is because what content a concept has—in virtue of which conceptual disputes arise—is, to some extent, up to concept users.[13] The conceptual engineering approach assumes that there are many possibilities on what the content of a given concept could be. The approach aims to improve the methods of approaching conceptual disputes.

Example of Free Will

Conceptual disputes are common in philosophy. The debate about FREE WILL is a case in point: compatibilists (e.g. Fischer & Ravizza, 1998; Frankfurt, 2003; Wolf, 1990) offer incompatibilists (e.g. Kane, 1998; van Invagen, 1983) different ways of making FREE WILL precise (and *vice versa*). Both sides defend certain conceptual choices. Each side offers arguments for certain conceptions of FREE WILL. Such philosophical work results in *maps* of conceptual possibilities: ways of

[11] This draws on Rawls' (1999: 5) terminology, though it is unclear whether he has the same distinction in mind.

[12] For a better understanding of what "grounding" *of* disagreements means in this context, see Chalmers (2011). Whether a given dispute meets the condition for a conceptual dispute is hard to establish conclusively. In our discussion, we rely on the following inference to the best explanation: When the literature on a dispute over a question has already answered the question in different ways, and when the given answers differ in their respective assumptions about the content of underlying concepts, the best explanation for the persistence of the dispute is that the dispute is at least partially grounded in disagreements over the content of some underlying concepts.

[13] Strictly speaking the assumption is that concept users can influence conceptual content. Whether this assumption is true and how such choices are made in practice are questions far beyond the scope of this paper. This paper explores what a conceptual engineering approach in AI responsibility would look like on the assumption that this approach is plausible.

(systematically) making a concept's content precise—within the realm of what we can recognize as possible precisifications of the concept.

Uncovering conceptual possibilities is useful and important work—hence our view that conceptual disputes are opportunities for philosophical progress. Once the conceptual terrain has been mapped so that conceptual choice points become clear, the philosophical question in dispute can be answered—and, in this sense, *given a certain understanding of the underlying concepts*, its philosophical problem "solved." For example, it might be true that free will *is* compatible with determinism, *if* FREE WILL is understood along the lines that compatibilists suggest. But, it may also be true that free will is *not* compatible with determinism, *if* FREE WILL is understood along the lines incompatibilists suggest. That the dispute persists, despite the availability of such maps indicates that the dispute is grounded, at least to a significant extent, in a disagreement about the content of FREE WILL; in other words: the dispute is a conceptual dispute.[14]

When a dispute is recognized as a conceptual dispute, the debate should attend to what conceptual choices would be *correct* (we will return later to the question as to what it could and should mean for conceptual choices to be correct). Given that whether free will is compatible with determinism depends on how FREE WILL is understood—how *should* FREE WILL be understood? Given that our actual understanding of "free will" is not how FREE WILL is necessarily to be understood, what way, if any, is the correct way of understanding FREE WILL?

Responsibility Gap Dispute as a Conceptual Dispute

With this picture of conceptual disputes in place, it becomes clear that the dispute about the AI responsibility gap is a conceptual dispute. Whether there are AI responsibility gaps depends on several crucial choices about the content of RESPONSIBILITY.

The diagram in Figure 1 is an example of a map of conceptual possibilities. It projects the terrain of conceptual choice points as a flow chart. This diagram illustrates how answers to conceptual questions about RESPONSIBILITY set one on a path that leads to or away from an AI responsibility gap. For example, take the question of whether RESPONSIBILITY presupposes

[14] This is a substantive meta-philosophical view that we cannot defend here at length. It contrasts with a deflationist view. A deflationist view would say that once there are conceptual maps, the philosophical work is done and the philosophical problem is "solved". If any questions remain, they concern the extension of concepts—e.g. whether there is FREE WILL in this world. We, by contrast, contend that philosophical disputes persist for good reason and the philosophical work is not exhausted by the creation of such maps.

control. If one answers this in the negative, and one assumes that an AI's human operator meets all other relevant necessary conditions for responsibility, the responsibility gap is avoided.

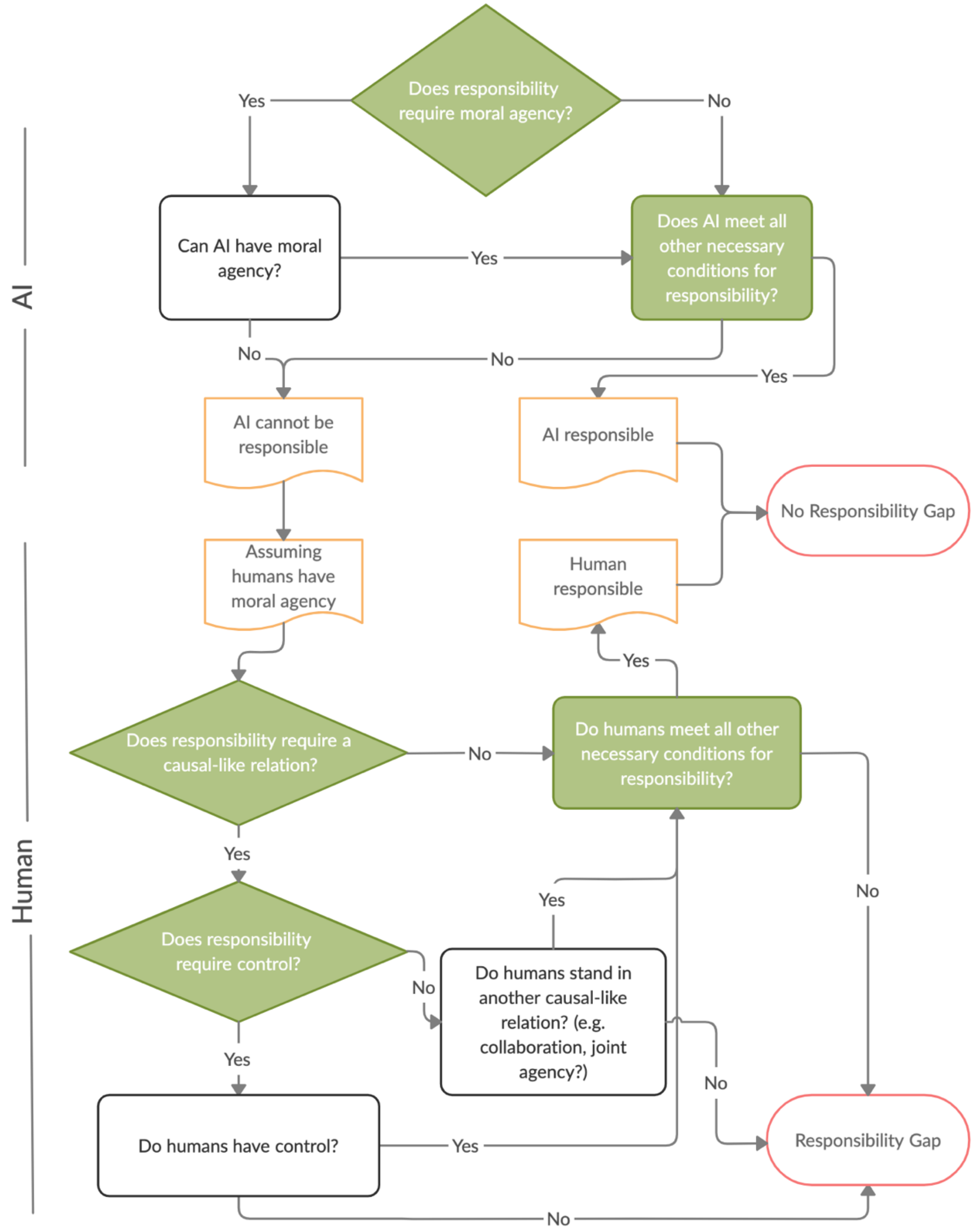
Does responsibility require moral agency?
Yes
No
Can AI have moral agency?
Yes
Does AI meet all other necessary conditions for responsibility?
No
No
Yes
AI cannot be responsible
AI responsible
No Responsibility Gap
Assuming humans have moral agency
Human responsible
Yes
Does responsibility require a causal-like relation?
No
Do humans meet all other necessary conditions for responsibility?
Yes
Yes
No
Does responsibility require control?
No
Do humans stand in another causal-like relation? (e.g. collaboration, joint agency?)
Yes
No
Do humans have control?
Yes
Responsibility Gap
No
AI
Human

Figure 1: A map of conceptual choice points for the responsibility gap problem focused on RESPONSIBILITY (in green). Diamond shapes indicate conceptual questions, i.e. questions about the *intension* of a concept. Square shapes indicate questions about the *extension* of a concept. Wave-squares indicate intermediate conclusions or assumptions.

Conceptual maps, such as the one in Figure 1, can demonstrate that a dispute is a conceptual dispute. A dispute is conceptual if all associated flow diagrams are such that all paths that lead to answers to the main question pass through at least one conceptual question.[15] This is one way of understanding the idea that the dispute over whether there is a responsibility gap is grounded in conceptual questions.

The conceptual map in Figure 1 is simplified. It represents only some of the conceptual questions in the responsibility gap dispute. For the purposes here, we concentrate on "AI" and "Human" as possible agents (see the labels on the left in Figure 1). Moreover, we concentrate on whether RESPONSIBILITY requires control or some other causal-like relation (and the related question about the extension of CONTROL: whether humans have control or whether they stand in some other causal-like relation to AI, such as collaboration or joint agency). We set aside other necessary conditions for RESPONSIBILITY, such the epistemic or the intentional condition.

But already this simplified conceptual map offers guidance. It identifies on what conceptual choices it depends whether there is a responsibility gap for AI.

Notice that the first question at the top of the diagram is about the meaning of RESPONSIBILITY, or what we can call the *intension* of the concept. In Figure 1, such questions are indicated by diamond shapes. The second question, by contrast—whether AI *has* moral agency—is about the *extension* of a concept, that is, whether the concept AGENCY extends to AI. Whereas the first question is about the *content* of RESPONSIBILITY, the second question is a question about *whether* AI falls under that concept.

The diagram highlights that questions about responsibility (the property) are not only about RESPONSIBILITY (the concept). The intension of RESPONSIBILITY may include other concepts, for example AGENCY or CONTROL. Thus, the extension of RESPONSIBILITY then depends on the extension of these other concepts. In this way, questions about the extension of a concept— in the diagram in each square—masks further conceptual questions and questions about the extension of other concepts. Each square could be unpacked into further diamonds and further squares.

Contributions to an existing literature can be projected onto conceptual maps. Some people answer both questions at the top—whether RESPONSIBILITY requires AGENCY and whether AI falls under AGENCY—in the affirmative. For example, Floridi and Sanders (2004) can be read as contending that AI *can* have moral agency. Others, by contrast, hold a different conception of RESPONSIBILITY: they deny that RESPONSIBILITY requires agency. Hellström

[15] The same dispute can be associated with slightly different diagrams. The complexity of presentation is a matter of expositional choice. Specifically, choices about the extensions of related concepts (such as CONTROL) can be unpacked into choices about extensions and intensions of this related concept (e.g. whether CONTROL requires an absence of intervening agency).

(2012) contends that instead a new concept of AUTONOMOUS POWER is needed. In Figure 1, Hellström's position would answer "No" to the first question but then "Yes" to the subsequent question. Either way—whether because AI have agency (as Floridi might argue), or whether they are in the extension of some other concept that is necessary for RESPONSIBILITY (as Hellström might argue)—we are on our way to the conclusion that AI can be responsible and that, therefore, a responsibility gap can be avoided.

Turning to the responsibility of humans (in bottom part of Figure 1), the central conceptual questions are whether RESPONSIBILITY requires control or some other causal-like relation, and whether humans can stand in this relation. All proponents of the responsibility gap problem assume that RESPONSIBILITY requires some kind or degree of control, and they argue that AI somehow undermines control (e.g., Danaher 2016, Matthias 2004, and Sparrow 2007). Matthias (2004: 177), for example, writes that "nobody has *enough control* over the machine's actions to be able to assume the responsibility for them". Similarly, Sparrow (2007: 71) argues that commanders are *not* responsible by *reductio*: if they were responsible for what AWS do, then "[m]ilitary personnel will be held responsible for the actions of machines whose *decisions they did not control*". In sum, if RESPONSIBILITY requires control and if a conception of CONTROL is adopted on which control either cannot be had over AI or is undermined by AI, then responsibility gaps seem unavoidable.

But the diagram also makes clear the paths on which responsibility gaps are avoided via human responsibility. One option assumes that RESPONSIBILITY requires control and that humans have the relevant sort of control (see e.g. Himmelreich, 2019; Simpson & Müller, 2016). Another option assumes that RESPONSIBILITY requires some other causal-like relation—such as supervision or collaboration—instead of control and contends that humans stand in this relation (see e.g. Köhler, 2020; Nyholm, 2018; Robillard, 2018).[16] If responsibility travels along the lines of supervision, then responsibility gaps can be avoided.

The picture that emerges is this: Regardless of how the existing literature answers the question of whether there are AI responsibility gaps, it often does so by—implicitly or explicitly—making choices on underlying conceptual questions. The literature on the responsibility gap has explored the different conceptual alternatives to good measure. Given, then, that the dispute about AI responsibility gaps is a mature conceptual dispute: Where to go from here?

[16] Whether supervision is a form of control or another form of causal-relation short of control depends on conceptual questions about CONTROL.

3. The Way Forward: Responsible AI as a Conceptual Engineering Problem

Once the terrain of a conceptual dispute has been sufficiently mapped so that the conceptual choice points are roughly understood, we can turn to the question of which choices would be *correct*: what conception is the *correct* one for the concept?[17] A crucial question is now what making the "correct" choices means here.

It is natural to think that finding the correct choice amongst the different conceptual choice-points is just figuring out what the *actual* content of our concept RESPONSIBILITY is, by engaging in *conceptual analysis*. One very plausible view as to what we should do when we engage in conceptual analysis comes from Jackson (1998: 30-37). On this view, conceptual analysis tries to determine our *folk-theory* of a concept, by considering what content would make best sense of our dispositions to apply it. Conceptual analysis, on this view, proceeds via the method of reflective equilibrium, by determining what conception makes most sense of our intuitions associated with the concept. Of course, such a conception can be mildly revisionary—and it should be, as it is unlikely that any coherent conception can preserve *all* of our intuitions (a point nicely made by Allan Gibbard (1992: 32) and which Jackson (1998: 35/36) himself highlights). However, the main aim of conceptual analysis is to preserve and make sense of the intuitions that we have. On this first view of what it means to make the "correct" conceptual choices, correct conceptual choices are revealed through conceptual analysis.

We think that this is *not* the way forward. The *actual* content of concepts lacks the relevant normative significance to conclusively answer whether there can be AI responsibility gaps—or any other normative question for that matter. *Even if* we found evidence for what the *actual* content of our concept of RESPONSIBILITY is, it may yet not be the *best* content, the one that *ought* to be associated with the concept. Given the myriad alternatives that could precisify the concept, privileging conceptual content just because it encodes a folk theory seems unwarranted. This is especially so, if it is possible to instead abandon our actual conception of RESPONSIBILITY in favor of one that evades certain problems, such as, for example, the creation of responsibility gaps. We could, and perhaps should, be more revisionary about the content of our concepts. At the very least, we should engage in a *normative* assessment of our concept and consider what conceptions suit important purposes, rather than look for the conception that fits best with our current disposition to apply it.

[17] An on-going more detailed mapping of the available and alternative conceptual possibilities will also be important.

Thus, rather than trying to resolve conceptual disputes by finding out what our actual conception of a concept is, they should be resolved through *conceptual engineering*. To clarify what this means, let us first explain what conceptual engineering is.

Conceptual Engineering

Conceptual engineering aims to systematically *evaluate* and *improve* our conceptual repertoire in some way or other. Conceptual engineering, thereby, explicitly wants us to take a stance that goes *beyond* trying to find the content that makes best sense of our dispositions to apply our concepts. Instead, conceptual engineering brings into focus what reasons we have to use some concepts in the first place and, maybe, to change what concepts we use, how we use them or what conceptions to associate with them. As we will understand it, conceptual engineering is a methodological approach that is concerned with *concepts*, as well as the *words* used to express them.

To some extent, conceptual engineering has, plausibly, always been a central business of philosophy, just not under that name (see e.g. the examples in Cappelen, 2018: 9-27). In fact, its core methodological ideas have been championed before (e.g. Bishop, 1992; Carnap, 1950; Haslanger, 2012). Recently, the topic is being discussed widely and investigated systematically (Burgess & Plunkett, 2013; Cappelen, 2018; Cappelen et al., 2020; Eklund, 2018; Plunkett, 2015).

How to understand conceptual engineering's core theses is a difficult question. Any answer depends on questions regarding the nature of concepts, meaning of linguistic expressions, and so on. Given our aims and limitations of space, we need not go into these further issues here since there is a growing literature on this (see e.g. the discussion in Cappelen 2018 or the papers in Cappelen et al., 2020 for an introduction). We understand the core suggestion of conceptual engineering as follows: A systematic investigation is possible and desirable of what conception *ought* to be associated with a concept.

Once we have a conceptual map of a dispute, the central question of a conceptual dispute is what conceptual choices—or what conception overall—would be *correct*. On the conceptual engineering approach, the *correct* conception is the one that we *ought* to associate with the concept. Hence, approaching this question with conceptual engineering means investigating the normative significance of the concepts involved: it is to engage in *normative* inquiry with regards which of the possible conceptions *ought* to give the content of the concept. Any particular conception must explain why we should make these particular conceptual choices, *rather than others*.

Such arguments will, hence, be specifically addressed to the question why we *ought to* use that particular conception and, so, side-step the worries we raised for conceptual analysis.

Engineering DEMOCRACY

An example might help to further clarify this. Take the word "democracy" and the associated concept DEMOCRACY. Let's assume, for the sake of illustration, that DEMOCRACY represents *majoritarian group-decision-making*, so that in its extension fall all cases where a group decides in accordance with what most of its members want. In this case, there is a determinate *conception* associated with the concept DEMOCRACY. Clearly, though, several other *possible* conceptions of the same concept exist. This raises a question: even if *majoritarian group-decision-making* was the conception that made most sense of our applications of DEMOCRACY, is that the conception that *ought* to give its content?

One way to look at this question is to consider what DEMOCRACY does for us. For example, both the word "democracy" and the concept DEMOCRACY have significance beyond what they represent. Specifically, they have a certain justificatory flavor: Democratic decisions are presumed to have a certain sort of *moral desirability* and *legitimacy*. Moreover, DEMOCRACY plays a certain role in our cognitive economy: We are inclined to defer to and respect decisions made democratically. When we deliberate about how to make group-decisions we assign special weight to democratic rule as a decision procedure. Given the important practical and theoretical role that DEMOCRACY plays for us, we should consider: What is the conception that gives a content for DEMOCRACY that fits best with its normative significance? Which conception, really, is *best* given the importance of DEMOCRACY for our lives?

Another way to look at conceptual engineering is to consider what DEMOCRACY *could* do for us. For example, we can engineer expressions or concepts to highlight problems, such that these problems can be rectified. Famously, Haslanger (2000) argues that "woman" should be associated with a conception of WOMAN that refers to the member of an oppressed social group. What it means to be a woman, on this approach, is to be oppressed. As such, the concept is associated with a problem that needs to be rectified.

This same ameliorative strategy could be used for DEMOCRACY: "democracy" could refer to a conception that is associated with a problem in order to highlight or identify shortcomings. For example, the conception of DEMOCRACY might require that a society is socially and economically egalitarian. On this conception, a nation that incarcerates a significant part of its population and that has sustained vast inequality of wealth is not a democracy. The US would, then, *not* be a democracy. Saying "the US is not a democracy", given the justificatory flavor of "democracy" generally, makes pragmatically clear that something is amiss.

Taking a step back, this discussion highlights two important points. First, both words and concepts *do* certain things for us—interests are at stake when words or concepts are used. These interests are quite varied: For some words or concepts we might be interested in *carving nature at its joints*, or *facilitating our understanding about reality*. But not all our interests are representational in this way. As our example shows, e.g. we have *practical* interests that DEMOCRACY serves or can serve. Second, conceptual engineering raises important *normative* questions. For one, we can ask what interests are most important. This is the question about the *function* the concept *ought* to perform. Moreover, we can ask which conception best satisfies these interests. These two questions are questions for conceptual engineering. These are the two questions that we focus on here.

Responsibility Gaps as a Conceptual Engineering Problem

We can now put our argument together: The dispute about whether there is an AI responsibility gap is grounded to a significant extent in conceptual questions. It is a conceptual dispute. The dispute is rooted in a conflict of different conceptions of RESPONSIBILITY. The question "is there a responsibility gap for AI?" hence comes down to the question: What is the *correct* conception of RESPONSIBILITY? The best way of approaching *this* question is through conceptual engineering. Specifically, we should understand this question as the question "What conception *ought* to give the content of RESPONSIBILITY?". The responsibility gap problem should therefore be seen and approached as a conceptual engineering problem.

The following questions likely arise with regards to this suggestion: How can a normative dispute over underlying responsibility conceptions proceed? How should we make and evaluate choices between alternative conceptions? What are the terms of such debate? — These are good methodological questions that concern both what determines what conception ought to give the content of a concept and how we find out about this. Each of these questions is debated in the evolving literature on conceptual engineering. We empathically welcome discussion about them specifically for the purposes of addressing the responsibility gap problem. This is the way to move the discussion forward. But we cannot discuss these matters here in full. Instead, what we will do in the rest of the paper is to kickstart further debate about these questions: We make a suggestion on how to go about engineering RESPONSIBILITY, where the argumentative burdens will lie if we take this road, and what avenues for engineering

RESPONSIBILITY are *prima facie* promising.[18] On the approach that we suggest, what the *correct* conception for RESPONSIBILITY is, should be determined by the important functions RESPONSIBILITY plays and what conceptions would allow it to perform that function best.[19] We demonstrate what conceptual engineering on this suggestion looks like, by first suggesting a list of functions that RESPONSIBILITY, plausibly, plays, and then using these to evaluate and argue for conceptions that avoid responsibility gaps.

4. Groundwork for Engineering RESPONSIBILITY in the Age of AI

Let us start with some clarifying remarks. First, on our view, RESPONSIBILITY is the concept we use to regulate certain kinds of emotional and practical responses and practices, namely those that we associate with holding one another morally responsible. For example, when people judge someone to be morally responsible for a harm, they will be inclined to blame them for this harm. People who find themselves morally responsible for a harm will be inclined to offer reparations for that harm in order to find forgiveness. When we find someone's capacities for action impaired, we are inclined to excuse them from blame, even if we find them responsible. And so on. RESPONSIBILITY is the concept these responses and practices are structured around.[20]

Second, RESPONSIBILITY, as many concepts, is indeterminate. Conceptual choices will yield a "responsibility-conception", that is, a complete and determinate conceptual content that the relevant responses and practices could *feasibly* be structured around. It seems plausible that there are many different responsibility-conceptions.

When we engage in conceptual engineering for RESPONSIBILITY, the central question is this: Which of the many different possible responsibility-conceptions ought to regulate our

[18] There is already one extensive discussion of RESPONSIBILITY as a conceptual engineering problem by Manuel Vargas (2013). However, Vargas is only concerned with the case of *human* responsibility as it figures in the classic free will debate, not with the special problem posed by AI.

[19] In using this functional approach, we follow proponents of conceptual engineering like Amie Thomasson (2020, 2021), Sally Haslanger (2000) or Mona Simion and Christoph Kelp (2020). Of course, not everyone agrees that this is the best approach (e.g. Cappelen, 2018), but going into this discussion would go beyond the paper's purposes and detract from our illustrative aims.

[20] The close connection between RESPONSIBILITY and these practices has first been noted by Peter Strawson (1962), who held the view that to be responsible just *is* to be a fitting target of these practices. Of course, this suggestion raises the question what FITTING means here. This is itself a conceptual engineering problem for RESPONSIBILITY (for different suggestions deviating from Strawson's conception, see e.g. Rosen, 2015; Wallace, 1994; Zimmerman, 2015).

responsibility-related practices? That is: which of these possible conceptions ought to give the content of RESPONSIBILITY? As we've suggested above, an adequate answer to *this* question needs to determine what our *most important* interests are when it comes to employing RESPONSIBILITY and what conception of RESPONSIBILITY fits these best (or at least sufficiently well).

We suggest to think about these interests as the *functions* that RESPONSIBILITY plays or ought to play. A concept's function, as we understand it, captures one way in which the concept figures in cognition or in practical or theoretical reasoning. For example, the concept of GREED, plays at least two functions. When you describe someone as "greedy", this, first, implies something about their behavior and, second, it also implies a moral evaluation of them as a person. Describing a concept as having a function does not mean, of course, that this concept has this function only if it plays this role in cognition and reasoning without exception all the time for all concept users. Instead, it does mean that our responsibility-related practices would be *missing* something important if we lacked the concept or that function.

To address the *responsibility gap problem* using conceptual engineering, we need to consider how well responsibility-conceptions that *close* responsibility-gaps perform, compared to possible responsibility-conceptions on which there *are* responsibility-gaps. Of course, we already have seen why responsibility gaps are problematic (see Section 1). In this sense, we already have identified some functions of RESPONSIBILITY, namely, those that, if they fail, make responsibility gaps problematic. But the questions have been left open what it is about RESPONSIBILITY that avoids these problems as well as what other functions RESPONSIBILITY should play. We will now briefly highlight some of the most important of these functions, to then indicate how this can inform strategies to engineer RESPONSIBILITY such that responsibility gaps are avoided.

Functions of RESPONSIBILITY

Let us first clarify how we determine the functions that a concept might play. Roughly, functions are identified by considering what using the concept allows concepts users to do, by identifying aims, purposes, or interests that are served by the use of the concept. As should be clear, there are, consequently, many ways to identify functions. Our investigation will be guided by suggestions that have shaped different philosophical approaches to RESPONSIBILITY. We assume that this is feasible, because philosophical reflection about conceptions is often guided, implicitly or explicitly, by deliberation about what it is concept users might be *doing* with the concept.

So, what are the functions that RESPONSIBILITY plays? One set of interests here is clearly practical, namely the interests that make the set of emotional responses and practices that we associate with moral responsibility normatively significant. Another set of interests here is theoretical: There might be interesting kinds in reality that the concept can pick out. However, for RESPONSIBILITY the practical dimension is clearly much more important, so our focus will be there.

There are at least six noteworthy such practical functions. Of course, we do not want to suggest that these carve up the full terrain of possible functions or that identifying the functions in this way is the best. The list below is meant to be illustrative not exhaustive. At best, the list should be taken as a first proposal that further research on how to engineer RESPONSIBILITY can build on.

First, RESPONSIBILITY plays *a ledger function*:[21] Responsibility-concepts can be used to facilitate a form of moral accounting, to keep track of what can be attributed to whom. The ledger is a metaphor for an overall assessment and record of a person's conduct. RESPONSIBILITY, on this function, grounds an evaluative assessment of someone on the basis of their actions, mental states, or events connected to them in certain ways. RESPONSIBILITY in this function is what allows us to say that someone is *cruel* in virtue of the intentions behind their actions. Note that this function for RESPONSIBILITY is entirely backward-looking, as it is only used to connect what happened for the assessment of a person. Furthermore, on this function, part of the content of RESPONSIBILITY will be further evaluative concepts, because to attribute responsibility *is* to make an evaluative assessment of a person's conduct. The accounting is, hence, grounded in part in a normative theory (at least the theory underlying the person's evaluative judgments). Such a ledger function of RESPONSIBILITY is highlighted by the views of, e.g., Gideon Rosen (2015) Gary Watson (1996) or Michael Zimmermann (1988, 2015).

Second, RESPONSIBILITY plays an *answerability function*. One important part of our responsibility practices is to determine who must be able to provide both explanatory and justificatory reasons as to why something happened and how to relate to those who must answer, if no such reasons are forthcoming. For example, when a bridge collapses, we need to identify who has to explain why it collapsed, whether there is anyone who owes excusing or justificatory reasons for the collapse and we must determine what to do if someone who owes such reasons fails to provide them. Furthermore, it is also a dimension of this part of the practice to single out those who must make sure that e.g. bad things do not happen again. The answerability function of RESPONSIBILITY corresponds to the interest we have in this part of that practice.

[21] The term "ledger view" is due to Fischer and Ravizza (1998) 8-9.

Note that in this function RESPONSIBILITY also enables us to keep books on what can be attributed to whom in some sense. The difference to the ledger function, however, is that the answerability function of RESPONSIBILITY is not to evaluate people. Rather, it allows us keeps track of who *must answer* for certain outcomes or events. These come apart. For example, parents might have to answer for what their small children do and might be required to apologize if they cannot offer good reasons for their children's behavior, without it yet being the case that the *parents* are evaluated based on their children's behavior. The answerability function prominently shapes the views of e.g., T.M. Scanlon (2008) and Angela Smith (2012, 2015).

Third, RESPONSIBILITY also plays a *communicative-educational function*. Tagging someone as responsible for something can serve to communicate moral expectations and build moral community by signaling that we "see them as individuals who are capable of understanding and living up to the norms that make for moral community" (McGeer, 2010: 303). Similarly, we use RESPONSIBILITY in the process of bringing individuals (such as children) into the moral community as participating members and in shaping how others respond to reasons. At the same time, RESPONSIBILITY used in this way can be used to express, communicate, and educate about the moral norms at play in the moral community, provide the necessary "scaffolding" of each other that is required for moral community, and to enable collective deliberation about the norms that shape the moral community. Importantly, RESPONSIBILITY plays this role not just through being communicated to others, but also by structuring our deliberation and guiding our responses in certain ways. Without RESPONSIBILITY and its associated responses and practices we would miss an important instrument for facilitating a central form of social communication, education, shared deliberation and community building. The *communicative-educational function* corresponds to this need. Note that when it plays this function, RESPONSIBILITY is closely connected to acceptance of certain moral norms. So, as with the ledger function, when it plays the communicative-educational function, the content of RESPONSIBILITY has further moral concepts build into it. Victoria McGeer (e.g. 2010) and Manuel Vargas (e.g. 2013) assign this function prominent importance.

Fourth, RESPONSIBILITY plays a *desert function*: RESPONSIBILITY can be and is employed to determine and keep track of what people *deserve*, or what treatment of them would be *just* or *fair*. This function of RESPONSIBILITY is especially visible in distributive and retributive *justice*. For example, *just* punishment for a wrong appears to belong to those and only those who are responsible for it. It also appears that someone *deserves* to be punished for an event only if they are responsible for its occurrence. And, of course, the same holds for reward, praise, and blame. Similarly, whether a distribution is just depends on whether those who have more (or less) have done something that makes them responsible for having more (or less). For example, if one person lost their house to a hurricane, whereas the other lost theirs to a game of poker, we would say that *ceteris paribus* the hurricane victim has a greater claim to be compensated for their loss. The reason is that, unlike the victim of the hurricane, the poker player is responsible for their loss.

In this use, RESPONSIBILITY serves to pick out something that partially grounds desert or that considerations about justice are sensitive to in a particular way. Specifically, RESPONSIBILITY functions to *relate* agents to actions, events, or outcomes in the right way, one that makes certain responses or reactions fitting with regards to demands of justice or desert. The desert function corresponds to our interest to treat people justly or fairly. It imposes an important *constraint* on our punishment and rewarding practices, as well as on our distributive regimes—such as taxes or college admission policies—and is a constraint that people are inclined to take very seriously. Whereas the ledger function is mostly about moral evaluation of agents, the desert function is concerned with the right treatment of agents.

When RESPONSIBILITY performs this function it, again, comes with strong relations to other normative concepts, specifically concepts such as DESERT, JUSTICE, or FAIRNESS. The desert function has played a prominent role in the free will debate, as it is often thought that RESPONSIBILITY in this sense requires free will, though, of course, that debate disagrees as to what the desert function actually requires.

A fifth function of RESPONSIBILITY is an *incentive function*: RESPONSIBILITY, through its attendant practices—such as praising and blaming—lays down incentive structures. Presumably, agents will typically avoid negative reactive attitudes like anger or resentment, as well as the kinds of behavioral modifications that follow blaming responses. Inversely, agents typically appreciate positive reactive attitudes like admiration and the behavior that follows. So, the reactive attitudes and the behavioral responses associated with responsibility can be a form of deterrence or attraction. This role of RESPONSIBILITY allows the concept to figure into agents' practical reasoning, as it offers at least prudential, if not moral, reasons for (or against) certain actions. Here, again, there will be a strong connection between RESPONSIBILITY and certain norms, namely those that using RESPONSIBILITY incentivizes adherence to. However, this connection need not be conceptual. The incentive function plays a central role on consequentialist conceptions of RESPONSIBILITY (e.g. Schlick, 1930; Smart, 1961), in analyses of responsibility drawing on criminal law (Duff, 2009), as well as in accounts of social norms (e.g. Brennan & Pettit, 2000)

Lastly, RESPONSIBILITY can play a *compensatory function*. This function is most tangible when the compensation takes monetary form. This function solves a social coordination problem: Who should compensate someone for the damages or injuries they suffered? That someone is responsible would mean, on this function, that they must pay for whatever damages they are responsible for—regardless of whether they caused the damage or whether what they did was morally wrong. The function of RESPONSIBILITY is to determine whether a party has to "make up" for some damage and who that party is. Suppose that during a storm, a tree in your neighbors' garden falls on your car, which now is severely damaged. Who—or whose insurance—is to pay for this? This can be a question that RESPONSIBILITY can answer, under its compensatory function.

The compensatory function hence resembles a function of the legal concept LIABILITY, which likewise seems to solve a coordination problem: Different countries treat the case of your damaged car differently. In some countries your neighbor is liable (as the owner of the tree) in other places, you are liable (as the owner of what has been damaged) and the accident is seen as an 'act of god'. Like LIABILITY, the compensatory function of RESPONSIBILITY might facilitate social cohesion and economic development. The compensatory function contrasts with the desert function, in that it operates instrumentally or pragmatically, whereas the desert function picks up on a property of a person that morally justifies holding them responsible. Unlike the communicative-educational function, the compensatory function aims to ensure only that damage or injury are compensated, it does not aim to communicate or educate. Nor does the compensatory function aim at regulating conduct; hence it differs from the incentive function. There is nothing that your neighbor should have done to prevent the tree from crashing on your car. Finally, the compensatory function differs from the answerability function. After all, as in the example of your neighbor's tree falling on your car, some cases require compensation without there being anything for the person responsible to answer for or to be educated about. Some argue that responsibility can work like strict liability to avoid responsibility gaps in AI (e.g. Floridi, 2017; Hevelke & Nida-Rümelin, 2015; Orly, 2014).

It is plausible that our current responsibility practices, perhaps depending on context and circumstances, are shaped by all of these functions. In fact, on closer inspection the functions are interconnected to some degree. However, the ability of RESPONSIBILITY to play these functions will differ across different possible responsibility-conceptions.

Avoiding Responsibility Gaps: Making Conceptual Choices based on Functions

We can now illustrate how the responsibility gap problem can be approached from a conceptual engineering perspective. Schematically put, there are broadly two starting points for conceptual engineering to avoid responsibility gaps. First, there is what we call a *function-first approach*. This approach starts by ordering the functions of RESPONSIBILITY that are most important to then investigates whether these functions entail conceptions of RESPONSIBILITY on which responsibility gaps arise. Second, there is what we call a *conception-first approach*. This approach starts with a responsibility-conception on which responsibility gaps do arise and asks whether there is an alternative responsibility-conception that performs the various functions of RESPONSIBILITY just as well as the original conception but on which responsibility gaps do not arise.[22] We now illustrate the conceptual engineering approach to AI responsibility on the conception-first approach.

Some responsibility-conceptions that generate AI responsibility gaps require a strong or demanding causal-like relation, such as control. For example, both Matthias (2003) as well as Sparrow (2007), contend that an AWS's commander is not responsible because they lack control. Yet, there are many conceptions of CONTROL. Matthias and Sparrow hold a responsibility-conception that assumes not only that RESPONSIBILITY requires a causal-like relation and that RESPONSIBILITY requires CONTROL, but that incorporates a conception of CONTROL on which those who operate an AI do *not* have sufficient control. We call such conceptions of CONTROL and the associated properties "strong" or "demanding." On responsibility-conceptions that incorporate such strong conceptions of CONTROL, there are responsibility gaps.

On a strong conception of CONTROL, human control might be undermined by the agency of AI. When the AI acts, e.g. against the plans of its operator or designer, the AI's agency makes it so that neither the operator nor designer might be able to prevent the outcome from occurring. On one such strong conception of CONTROL, "a system is under the control (in general) of an agent if, and to the extent to which, its behavior responds to the agent's plans, manoeuvres or operations" (Mecacci & Santoni de Sio, 2020: 105; the conception they describe is the one developed by John Michon (1985)). Strong conceptions of CONTROL have a particular kind of interventionist or causal flavor.

[22] This assumes a dominance criterium (the alternative conception is at least as good as the original conception in all respect and strictly better in at least one respect). A more complex picture of conceptual engineering involves trade-offs: When one conception is better than another in some respects but worse in others, is this conception preferable overall?

What, if any, conception of CONTROL is required for RESPONSIBILITY? Answering this question involves making an important conceptual choice. This choice can make the difference between responsibility-conceptions that generate responsibility gaps and those that do not. If RESPONSIBILITY requires a causal-like relation and if, more specifically, responsibility requires a certain interventionist conception of CONTROL, then there are responsibility gaps.

The conceptual engineering approach now invites the following question: Can we do better than a conception of RESPONSIBILITY that involves such a demanding conception of CONTROL and that, thereby, leads to responsibility gaps? Given that responsibility gaps are problematic (see Section 1), all other things being equal, it would be strictly preferable to have a responsibility-conception that does *not* lead to responsibility gaps. How we precisify concepts is to some extent up to us, the concept users—or so the conceptual engineering approach assumes—so we should explore whether there is a responsibility-conception that assumes either a weaker conception of CONTROL or some other less demanding relation, thereby avoids responsibility gaps, while fulfilling the functions of RESPONSIBILITY at least as well as the original responsibility-conception.

Alternative conceptions of RESPONSIBILITY that avoid responsibility gaps are available. One of those is the proposal by Nyholm (2018: 1217), who argues that "humans involved are responsible for what the robots do for the reason that they initiate, and then supervise and manage, these human–machine collaborations." We concentrate on this proposal here. Does this alternative conception fulfill the functions of RESPONSIBILITY just as well as the conception, held by Matthias and Sparrow, that assumes a strong conception of CONTROL?

Nyholm's (2018) conception of RESPONSIBILITY amounts to the following picture. First, an AI—such as AWS or AVs—has agency of a certain kind ("domain-specific principled supervised agency"). Because of this agency, it is true that operators and designers cannot "fully 'control and predict'" what the AI is going to do, as proponents of the responsibility gap argument contend (Nyholm 2018: 1205). However, "mere unpredictability and the inability to fully control a piece of technology do not by themselves appear to eliminate responsibility on the part of the user." (Nyholm 2018: 1206). In other words, full control—by which we assume from the context of the discussion Nyholm means a strong conception of CONTROL—is not necessary for RESPONSIBILITY. Rather, some other relation can be equally sufficient, at least in some relevant cases.

On Nyholm's conception of RESPONSIBILITY, the relation between the user or operator of an AI and the AI system itself is analogous to the relation between a parent and a small child. It is an on-going relationship of supervision. For example, in the case of AWS, "designers are paying close attention to whether the commanding officers are happy with the robot's performance. If not, the designers and engineers update the hardware and software so as to make its performance better track the commanding officers' preference and judgments" (Nyholm 2018: 1213). On this view, operators of an AI are responsible for behavior of the AI system because they supervise it, that is, because they maintain, improve, and teach the AI system what to do and how to behave. This relation of supervision also entails that supervisors have control in the sense that they can stop the AI system. A user of an AV, for example, "has the power to take over control or stop the car from doing what it's doing" (Nyholm 2018: 1209). However, having control in this sense is not sufficient for control in the strong sense. Thus, even if operators do not have control over an AI system according to a strong conception of CONTROL, so argues Nyholm (2018: 1214), they can be responsible. Hence, there are no responsibility gaps.

As should be clear, Nyholm's conception of RESPONSIBILITY is quite different from the one presupposed by Matthias or Sparrow. We can now systematically compare these conceptions. The conceptual engineering approach does this by transcending this first-order dispute about whether there is a responsibility gap. The conceptual engineering approach recognizes this dispute as conceptual and turns to the underlying conceptual questions. It asks: Which of the two conceptions is better? The conception by Nyholm is clearly better in one respect: there are no responsibility gaps. By identifying the functions of a concept, the conceptual engineering approach moreover offers a framework that guides the conceptual evaluation on which of the two conceptions is better overall. The conceptual engineering approach asks: Does Nyholm's alternative responsibility-conception still fulfill the functions of RESPONSIBILITY?

The procedure to answer this question is clear: Consider the relevant functions and examine whether assuming a strong conception of CONTROL or some other relation affects how well the resulting responsibility-conception fulfills the function.[23] Since the point here is only to illustrate the conceptual engineering approach for AI responsibility, we present this last step somewhat schematically. We discuss the ledger, answerability, compensation, communicative-educational, incentive, and desert functions in turn.

Consider the ledger function first. This function grounds an evaluative assessment of someone on the basis of their actions or events connected to them. This function obviously

[23] We consider this marginal change within a responsibility-conception instead examining each of the conceptions. We cannot give a full comparison because of space limitations but also because Matthias and Sparrow do not say enough about what they mean by "responsibility" and what functions they take RESPONSIBILITY to play.

requires *some* relation that links an agent with actions and events. But this relation need not be a strong conception of CONTROL, or even a causal-like relation. A different relation between an agent and actions, events, or occurrences can be sufficient to ground evaluative assessment of the agent. A harm that an AI causes can be attributed to the operator of the AI, following Nyholm (2018: 1214), because the operator collaborated with the AI and trained it. Thus, the operator might be evaluated on the basis of what the AI did. In fact, a responsibility-conception that incorporates such a relation might play the ledger function better than one that incorporates a strong one. This is because a person's conduct might not be under their strong control—think of unintentional omissions, or habits—but still be properly attributed to that person. A person might be properly morally evaluated on the basis of conduct that is not under their control, on the strong conception of CONTROL (see e.g. Shoemaker, 2015; Watson, 1996).

As for the answerability function: A responsibility-conception that incorporates a relation of supervision can play the answerability function just as well as—and in fact better than—one that incorporates a strong conception of CONTROL. As noted in the description of this function: The agent who must answer for an outcome can be different from the agent who caused, brought about, or controlled it. An agent can be answerable for some conduct even if they had little or no control—on the strong conception of CONTROL—over the outcome. Weaker relations allow this: they can accommodate that someone is answerable for the harm caused by an AI. Accordingly, some argue that as far as the answerability function of RESPONSIBILITY is concerned, even if supervisors have insufficient control, there is no responsibility gap (Burri 2017, Himmelreich 2019). In fact, a supervision relation might be ideally suited to ground the kinds of concerns related to the answerability function, as this sort of relation is already shaped by the norms and expectations we associate with answerability. After all, to be a supervisor for something or someone is in part already to be in a position of having to answer for the thing's or person's conduct, at least within the domain within which one is supervising. The causal relations required for supervision will, hence, fit the answerability function very well and, likely better than relations of strong control, because such relations are too narrow to include all plausible candidates for answerability.

The compensatory function similarly allows that the agent who must compensate for a harmful outcome can be different from the agent who caused or controlled this outcome. An agent may have to pay compensation even if they had little or no control over the outcome. This is because the compensatory function aims only to ensure that damages are compensated, and injuries are rectified. Its rationale, as explained above, is pragmatic and resembles that of strict liability. As such, to play the compensation function, CONTROL is not required. Assuming a weak instead of a strong conception of CONTROL therefore should not lessen how otherwise identical responsibility-conception can fulfill the compensatory function.

As for the communicative-educational function: A responsibility-conception like Nyholm's is particularly well-equipped to play the communicative-educational function. This is because the relation of supervision and collaboration between an AI user or designer and the AI system is largely one of education and training. If the AI system causes harm, the AI user or designer is the right agent to be held responsible for the communicative-educational function insofar as the AI user or designer can pass the communicated information on to the AI that they supervise and thereby "educate" the AI system.

Similar considerations apply for the incentive function. One might argue that without sufficient control, being held responsible cannot be an incentive because an agent would be unable to respond to this incentive with a change their behavior (or in behavior that they can influence). Therefore, to play the incentive function, a responsibility-conception needs to incorporate a strong conception of CONTROL. But this argument overlooks what Nyholm's argument brings out: The AI user *has* great influence over an AI insofar as only *they* train the AI system. Hence, holding the AI user responsible places the incentive correctly. Because an AI operator is best positioned to get the AI system to behave as the incentive requires, if RESPONSIBILITY should play an incentive function, the AI user should be responsible. A responsibility-conception that incorporates a weaker relation than that required by a strong conception of CONTROL therefore fulfills the incentive function well.

Finally, consider the desert function. For a person to deserve a certain responsibility response for an event or outcome, this event or outcome must be connected to this person in the right way. One might say that only a strong conception of CONTROL grounds someone's responsibility such that it allows for deserved blame and praise, reward and punishment. But this line of argument is too quick. First, weaker relations than strong control can, plausibly, play the role demanded by the desert function. Consider an example given by Nyholm (2018: 1212): an adult and a child are robbing a bank together, with the child doing most of the work and the adult initiating the robbery, but then staying mostly in the background. Suppose something goes seriously wrong and the child causes a grievous harm to one of the bystanders in the bank. Even though the adult lacks strong control over the child's behavior, it still seems plausible that the adult is blameworthy—they deserve to be blamed—for the harm that was caused by the child and that they deserve to be punished for the harm. Or, consider another example: assume that there is an indeterministic machine that will kill someone with a very small, but not negligible probability if you press a certain button. The machine has no other uses. Suppose someone presses the button and a person dies. Here, again, the relation that holds between action and outcome is not strong control, but it still seems that the person in question deserves blame and punishment. So, weaker relations than strong control can be enough—and may indeed be required—for RESPONSIBILITY to play the desert function. More importantly, a weak relation might at least be enough for our purposes in the case of AI use, given that these cases strongly resemble the kinds of cases just described.

Second, notice that in describing the desert function itself we are using concepts, such as DESERT or FAIRNESS. These concepts themselves, though, raise questions about what functions they serve and what sorts of conceptions would best serve them. What is it that the concept DESERT actually does for us and what *should* it do for us? What is noteworthy here is that conceptions of DESERT have already been put forward on which the point of DESERT is to facilitate some of the other functions of responsibility (see, in particular Vargas, 2013), for which we have already seen that a weaker relation than strong control is perfectly appropriate at least for the case of AI. What this shows us, at least, is that it is not obvious that the *best* conception for DESERT presupposes a strong conception of CONTROL. Furthermore, the general lesson to draw from these observations is that the conceptual engineering approach is infectious: When identifying the functions of one concept, e.g. RESPONSIBILITY, we must also assess the functions of related concepts, e.g. CONTROL, DESERT, FAIRNESS. So, even if a weaker relation than CONTROL might not fit the desert function, it is still perfectly possible that we should assess and revise our responsibility practices wholesale and adopt a less demanding conception of DESERT.

5. Conclusion

We have argued that AI does not raise responsibility problems—depending on how RESPONSIBILITY is understood. We argued that the literature on the responsibility gap problem is involved in a conceptual dispute. To make further progress, more attention should be given to the shape of the RESPONSIBILITY conceptions that are underlying the immediate—or first-order—question of whether there is a responsibility gap for AI.

We started from the basic premise that concepts are mutable and that different conceptions precisify their content along several choice points. By way of a cursory literature review, we described the conceptual choice points for the responsibility gap problem. We have then described the approach of conceptual engineering and we have sketched out how responsibility gaps can be investigated from this vantage point. Specifically, we have identified the different practical functions that RESPONSIBILITY may play. On the approach of conceptual engineering, these functions guide a systematic evaluation to improve our conceptual repertoire. On the resulting picture, that is if RESPONSIBILITY is engineered to fulfill the functions that we identified, responsibility gaps may not arise. We illustrated this approach by applying it to two conceptions of RESPONSIBILITY—only one of which gives rise to a responsibility gap. We argued that this conception is a better conception.

The approach of conceptual engineering may seem deflationary or disappointing. It may appear to give a somewhat unsatisfactory answer to the question of whether there are responsibility gaps for AI: It depends. You can engineer responsibility-conceptions in different ways—on some responsibility gaps may arise, on others not.

But a more hopeful outlook is that this "it depends" answer is exactly what philosophical progress looks like. Conceptual engineering makes good on the platitude that philosophy helps to understand questions better, even if it does not settle them. To do so, conceptual engineering moves the attention to higher-order questions about the concepts involved—what functions they should fulfill and what interests they should serve—and related methodological questions of how disputes over conceptual content ought to be conducted.

Moreover, conceptual engineering finds concrete ways out of the AI responsibility gap problem. As we have sketched in this paper, there might be conceptions of RESPONSIBILITY that fulfill all functions of RESPONSIBILITY without allowing for responsibility gaps. And, for those responsibility-conceptions that give raise to responsibility gaps, the approach points us towards ways of improving them.

Engineering responsible AI is hence an undertaking on two fronts. First, engineering responsible AI is a practical engineering task. Robots need to be built; software needs to be developed. All this needs to be done in a way that the AI–user nexus meets the requirements of a responsibility-conception. Second, engineering responsible AI is also a theoretical engineering task. Concepts are "up to us". In the vein of the recent literature on conceptual engineering, we propose that responsible AI can be engineered through a deliberate choice of responsibility-conceptions. The central disputes around which the responsibility gap literature has revolved so far—whether RESPONSIBILITY requires control, whether operators of AI have control in the right sense—can be resolved by describing the conceptual desiderata and trying to systematically improve our conceptual repertoire.

Responsibility gaps hence need to be closed from at least two ends: On the one end, practical engineering and deployment practices need to be appropriate such that they can be called "responsible". On the other end, the responsibility-conceptions might need to evolve or be deliberately adapted to foreclose the possibility of responsibility gaps while ensuring that RESPONSIBILITY plays the roles that it ought to play.

6. References

Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a Framework for Levels of Robot Autonomy in Human-robot Interaction. *J. Hum.-Robot Interact.*, *3*(2), 74–99. https://doi.org/10.5898/JHRI.3.2.Beer

Bishop, M. A. (1992). The Possibility of Conceptual Clarity in Philosophy. *American Philosophical Quarterly*, *29*(3), 267–277. http://www.jstor.org/stable/20014420

Braham, M., & van Hees, M. (2010). Responsibility Voids. *The Philosophical Quarterly*, *61*, 6–15.

Brennan, G., & Pettit, P. (2000). The Hidden Economy of Esteem. *Economics and Philosophy*, *16*, 77–98.

Bullock, J. B. (2019). Artificial Intelligence, Discretion, and Bureaucracy. *The American Review of Public Administration*, *49*(7), 751–761. https://doi.org/10.1177/0275074019856123

Burgess, A., & Plunkett, D. (2013). Conceptual ethics I & II. *Philosophy Compass*, *8*(12), 1091–1011 and 1102–1110. https://doi.org/10.1111/phc3.12085

Burri, S. (2017). What's the Moral Problem with Killer Robots? In R. Jenkins, M. Robillard, & B. J. Strawser (Eds.), *Who Should Die?* Oxford University Press.

Campaign to Stop Killer Robots. (2021). *The Problem*. Https://Www.Stopkillerrobots.Org.

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

Cappelen, H., Plunkett, D., & Burgess, A. (Eds.). (2020). *Conceptual Engineering and Conceptual Ethics*. Oxford University Press.

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press.

Chalmers, D. J. (2011). Verbal disputes. *Philosophical Review*, *120*(4), 515–566. https://doi.org/10.1215/00318108-1334478

Coeckelbergh, M. (2016). Responsibility and the Moral Phenomenology of using Self-Driving Cars. *Applied Artificial Intelligence*, *30*, 748–757.

Collins, S. (2019). Collective Responsibility Gaps. *Journal of Business Ethics*, *154*, 943–954.

Danaher, J. (2016). Robots, Law and the Retribution Gap. *Ethics and Information Technology*, *18*, 299–309.

de Sio, F. S., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers Robotics AI*, *5*(FEB), 1–14. https://doi.org/10.3389/frobt.2018.00015

Duff, R. A. (2009). Strict Responsibility, Moral and Criminal. *The Journal of Value Inquiry*, *43*, 295–313.

Duijf, H. (2018). Responsibility Voids and Cooperation. *Philosophy of the Social Sciences*, *48*, 434–460.

Eklund, M. (2018). *Choosing Normative Concepts*. Oxford University Press.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Floridi, L. (2017). *Roman Law Offers a Better Guide to Robot Rights than Sci-Fi*. Financial Times.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Machine Ethics and Robot Ethics*, 317–347. https://doi.org/10.4324/9781003074991-30

Fossa, F. (2018). Artificial moral agents: moral mentors or sensible tools? *Ethics and Information Technology*, *20*(2), 115–126. https://doi.org/10.1007/s10676-018-9451-y

Frankfurt, H. (2003). Freedom of the Will and a Concept of a Person. In *Free Will* (pp. 322–336). Oxford University Press.

Gibbard, A. (1992). *Wise Choices, Apt Feelings*. Harvard University Press.

Gunkel, D. (2017). Mind the Gap: Responsible Robotics and the Problem of Responsibility. *Ethics and Information Technology*.

Hakli, R., & Mäkelä, P. (2019). Moral Responsibility of Robots and Hybrid Agents. *The Monist*, *102*(2), 259–275. https://doi.org/10.1093/monist/onz009

Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be? *Nous*, *34*(1), 31–55. https://doi.org/m

Haslanger, S. (2012). *Resisting Reality*. Oxford University Press.

Hellström, T. (2013). On the Moral Responsibility of Military Robots. *Ethics and Information Technology*, *15*, 99–107.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Science and Engineering Ethics*, *21*, 619–630.

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology*, *16*(3), 197–206. https://doi.org/10.1007/s10676-014-9345-6

Himma, K. E. (2009). Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties must an Artificial Agent have to be a Moral Agent? *Ethics and Information Technology*, *11*, 19–29.

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

Hooker, J., & Kim, T. W. (2019). Truly autonomous machines are ethical. *AI Magazine*, *40*(4), 66–73. https://doi.org/10.1609/aimag.v40i4.2863

Jackson, F. (1998). *From Metaphysics to Ethics*. Oxford University Press.

Kane, R. (1998). *The Significance of Free Will*. Oxford University Press.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Köhler, S. (2020). Instrumental Robots. *Science and Engineering Ethics*, *26*, 3121–3141.

Köhler, S., Sauer, H., & Roughley, N. (2017). Technologically blurred accountability? Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Debiel (Eds.), *Moral Agency and the Politics of Responsibility* (pp. 51–67). Routledge.

Lechterman, T. M. (2022). The Concept of Accountability in AI Ethics and Governance. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudsin, A. Korinek, M. M. Young, & B. Zhang (Eds.), *Oxford Handbook of the Governance of AI*.

List, C., & Pettit, P. (2011). *Group Agency. The Possibility, Design, and Status of Corporate Agents*. Oxford University Press.

Liu, H. Y. (2017). Irresponsibilities, inequalities and injustice for autonomous vehicles. *Ethics and Information Technology*, *19*(3), 193–207. https://doi.org/10.1007/s10676-017-9436-2

Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, *6*, 175–183.

McGeer, V. (2010). Co-Reactive Attitudes and the Making of Moral Community. In R. Langdon & C. MacKenzie (Eds.), *Emotions, Imagination and Moral Reasoning*. Psychology Press.

Mecacci, G., & Santoni de Sio, F. (2020). Meaningful Human Control as Reason-Responsiveness: the Case of Dual-Mode Vehicles. *Ethics and Information Technology*, *22*, 103–115.

Michon, J. (1985). *Human behavior and traffic safety*. Springer US.

Nyholm, S. (2018). Attributing Agency to Automated Systems: On Human-Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, *24*, 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Orly, R. (2014). Don't Sue Me, I Was Just Lawfully Texting & Drunk When MY Autonomous Car Crashed Into You. *Southwestern Law Review*, *44*, 175–208.

Pagallo, U. (2011). Robots of Just War: A Legal Perspective. *Philosophy & Technology*, *24*, 307–323.

Plunkett, D. (2015). *Which Concepts Should We Use ?: Metalinguistic Negotiations and The Methodology of Philosophy. 58*, 828–874. https://doi.org/10.1080/0020174X.2015.1080184

Rawls, J. (1999). *A Theory of Justice. Revised Edition*. Belknap Press.

Robillard, M. (2018). No Such Thing as Killer Robots. *Journal of Applied Philosophy*, *35*, 705–717.

Roff, H. (2013). Killing in War: Responsibility, Liability, and Lethal Autonomous Robots. In F. Allhoff, N. Evans, & A. Henschke (Eds.), *Routledge Handbook of Ethics and War: Just War Theory in the 21st Century*. Rougtledge.

Rosen, G. (2015). The Alethic Conception of Moral Responsibility. In R. Clare, M. McKenna, & A. Smith (Eds.), *The Nature of Moral Responsibility. New Essays* (pp. 45–64). Oxford University Press.

Scanlon, T. M. (2008). *Moral Dimensions. Permissibility, Meaning, Blame*. Harvard University Press.

Schlick, M. (1930). *Fragen der Ethik*. Verlag Julius Springer.

Schulzke, M. (2013). Autonomous Weapons and Distributed Responsibility. *Philosophy & Technology*, *26*, 203–219.

Shoemaker, D. (2015). *Responsibility from the Margins*. Oxford University Press.

Simion, M., & Kelp, C. (2020). Conceptual Innovation, Function First. *Noûs*, *54*, 985–1002.

Simpson, T. W., & Müller, V. C. (2016). Just War and Robots' Killings. *Philosophical Quarterly*, *66*(263), 302–322. https://doi.org/10.1093/pq/pqv075

Smart, J. J. C. (1961). Free Will, Praise, and Blame. *Mind*, *70*, 291–306.

Smith, A. M. (2012). Attributability, Answerability, and Accountability: In Defense of a Unified Account*. *Ethics*, *122*(3), 575–589. https://doi.org/10.1086/664752

Smith, A. M. (2015). Responsibility as Answerability. *Inquiry*, *58*(2), 99–126. https://doi.org/10.1080/0020174X.2015.986851

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the Aristotelian Society*, *48*, 1–25.

Thomasson, A. (2020). A Pragmatic Method for Conceptual Ethics. In H. Cappelen, D. Plunkett, & A. Burgess (Eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford University Press.

Thomasson, A. (2021). Conceptual Engineering: When do we need it? How can we do it? *Inquiry*.

Tigard, D. W. (2020). There Is No Techno-Responsibility Gap. *Philosophy & Technology*.

Totschnig, W. (2020). Fully Autonomous AI. *Science and Engineering Ethics*, *26*(5), 2473–2485. https://doi.org/10.1007/s11948-020-00243-z

US Department of Defense. (2012). *Autonomy in Weapon Systems*.

van Invagen, P. (1983). *An Essay on Free Will.* Clarendon Press.

Vargas, M. (2013). *Building Better Beings. A Theory of Moral Responsibility*. Oxford University Press.

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI and Society*, *0123456789*. https://doi.org/10.1007/s00146-021-01189-x

Vladeck, D. C. (2014). Machines without principals: Liability rules and artificial intelligence. *Washington Law Review*, *89*(1), 117–150.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Harvard University Press.

Walzer, M. (1977). *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. Basic Books.

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics*, *24*(2), 227–248.

Wolf, S. (1990). *Freedom Within Reason*. Oxford University Press.

Zimmerman, M. (1988). *An Essay on Moral Responsibility*. Rowman and Littlefield.

Zimmerman, M. (2015). Varieties of Moral Responsibility. In R. Clarke, M. McKenna, & A. Smith (Eds.), *The Nature of Moral Responsibility* (pp. 45–64). Oxford University Press.

Competing interests

One of the authors held a postdoctoral fellowship funded by Apple Inc. researching the ethics of autonomous systems while spending time at Apple Inc. The authors declare that they have no other conflict of interest, or disclosures on declarations on competing interests, funding, ethical approval, or consent to publish.

Acknowledgements

This paper was presented at the AiTech Agora in Delft, the Robots and Responsibility Workshop organized by Leonhard Menges, and the FS Philosophy Forum. We are grateful for the discussions and suggestions we received at these venues. We thank Herman Veluwenkamp and Christine Tiefensee for extensive comments and discussions. S.K.'s research for this paper was conducted while he was a principal investigator of the project group Regulatory Theories of AI of the Centre Responsible Digitality (ZEVEDI).