

Dubito Ergo Sum: Exploring AI Ethics

Viktor Dörfler
University of Strathclyde Business School
Glasgow, United Kingdom
viktor.dorfler@strath.ac.uk

Giles Cuthbert
Chartered Banker Institute
Edinburgh, United Kingdom
giles.cuthbert@charteredbanker.com

Abstract

We paraphrase Descartes' famous dictum in the area of AI ethics where the "I doubt and therefore I am" is suggested as a necessary aspect of morality. Therefore AI, which cannot doubt itself, cannot possess moral agency. Of course, this is not the end of the story. We explore various aspects of the human mind that substantially differ from AI, which includes the sensory grounding of our knowing, the act of understanding, and the significance of being able to doubt ourselves. The foundation of our argument is the discipline of ethics, one of the oldest and largest knowledge projects of human history, yet, we seem only to be beginning to get a grasp of it. After a couple of thousand years of studying the ethics of humans, we (humans) arrived at a point where moral psychology suggests that our moral decisions are intuitive, and all the models from ethics become relevant only when we explain ourselves. This recognition has a major impact on what and how we can do regarding AI ethics. We do not offer a solution, we explore some ideas and leave the problem open, but we hope somewhat better understood than before our study.

Keywords: AI ethics, responsible AI, understanding, sensory knowledge, indwelling

1. Introduction

In this conceptual paper we argue for the ethical approach in AI that suggests leaving most of the moral issues in the hands of humans. This is not to say that we should not try to 'put' a moral perspective into AI, but that we also need to 'put' in the limitations. The natural first step is that we need to understand the limitations, but the far trickier next one is how to get AI to identify its limitations, and request human assistance. We do not intend here to get into the technical details of what can and needs to be done in AI, we remain at the level of philosophizing AI and the human mind in the context of ethics, thus problematizing AI ethics. We do this from a distinct phenomenological position, within a moderate interpretivist paradigm (Dörfler, 2023b).

In this paper we do not provide a generic review of the AI literature, we only explain the basic concepts that we use in this paper here in the introduction. Thus, for the purpose of this paper we use one of the oldest definitions of AI, back from the Dortmund days, according to which AI is loosely defined as machines that can accomplish tasks that humans would accomplish through thinking (e.g. Dörfler, 2020).

This definition does not say anything about AI accomplishing such tasks in a way that resembles human thinking; we do not see anything in this definition that implies that AI would think in the human sense of the word. Importantly, AI as a field is not simply a study of the machines, it is as much the study of the human mind (for a more detailed description see e.g. Dörfler, 2022; Dörfler, 2023a). Specifically in the area of decision-making we find Davenport's (2018, p. 44) description of AI as "analytics on steroids" particularly expressive and, consequently, AI does not make decisions but it can make our (human) decisions better informed.

Decision-making is an important aspect of using AI when it comes to ethics, and all (at least the vast majority of) our decisions have moral components. In this paper we do not engage with particular application areas of AI, such as medical diagnosis (Davenport & Glaser, 2022; Davenport & Glover, 2018; Göndöcs & Dörfler, 2023), we locate our interest loosely in organizations (Csaszar & Steinberger, 2022; Davenport & Euchner, 2023; Davenport & Miller, 2022; Glikson & Woolley, 2020; Grodal et al., 2023; von Krogh, 2018; Leavitt et al., 2021; Lindebaum & Ashraf, 2021), in which concept we include business organizations, government institutions, as well as organizations, such as hospitals and universities, regardless of whether they are for profit or not. We are conscious of the organizational learning aspects and implications of AI (Balasubramanian et al., 2020; Davenport & Ammanath, 2020; Davenport & Mittal, 2022, 2023; Göndöcs & Dörfler, 2022; Oliver et al., 2017; Pachidi et al., 2021; Raisch & Krakowski, 2021; Tschang & Almirall, 2021); although we cannot tackle these at a great depth here.

In order to develop our argument eloquently captured paraphrasing Descartes, in what follows, we begin with a brief but systematic overview of the most common approaches and models in the domain of ethics. This is followed by a review of the AI ethics literature. Then we outline our philosophical position and methodological considerations, before getting to the points we want to make. Each of the next three sections provides a component of our conceptual analysis. First, we explore the sensory grounding of knowledge, showing how such sensory grounding, in a sense we attribute it to humans, is not possible in AI. Second, we suggest that AI lacks understanding, and we illustrate this with recent events in the AI landscape. Third, based on the literature, we argue that doubting oneself is an essential ingredient of morality, and we show that this one requirement also incorporates the previous two. In our final commentary we discuss what can be done, and offer three points, to make the best use of AI in morally acceptable ways and indicate areas of further research.

2. The Vast Landscape of Ethics

In this paper the term ethics is used to designate a branch of philosophy, the discipline that studies morality. In turn, morality refers to dealing with the issues of good and evil, right and wrong, responsibility, and such. In this sense, ethics is one of the oldest topics studied by humankind. In the Western tradition we see philosophy starting with discussing metaphysics followed by epistemology, which is closely followed by ethics, which Socrates has brought centerstage. This means that we have at least some two and half millennia of literature to cover, therefore our review will not be comprehensive, but our review is systematic in the sense that the main philosophical models are organized into categories (see Figure 1).

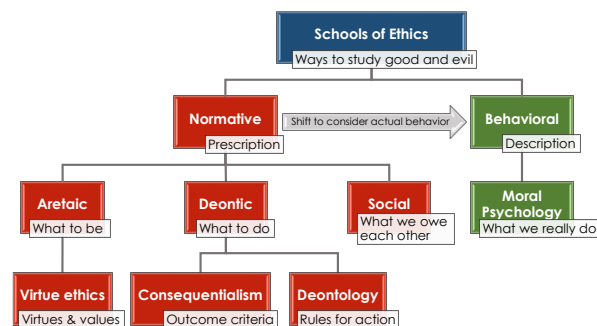


Figure 1. Overview of Historical Schools of Ethics

The normative schools attempt to prescribe what to be or what to do or not do; the behavioral school aims at describing what people actually do. The three dominant normative schools of ethics, are virtue ethics, rules-

based ethics (or deontology), and consequentialism. Other important models in normative ethics include pragmatist, intuitionist, contractualist, and feminist ethics. While these schools are all centered on individuals, there are also less known social variants, proposed by a small number of philosophers.

While it was Socrates who initially turned the philosophers' attention to considerations of morality, particularly to counter the sophist approach to teach anyone how to win debates using tools of rhetoric, it was Aristotle (cca 330 BCE) who composed ethics into the first systematic model called *virtue ethics*. Aretaic ethics, of which virtue ethics is the dominant example, focuses on features or characteristics that are desirable in individuals if they are to be considered to meet high moral standards; in virtue ethics these characteristics are the virtues. Importantly, virtue ethics does not aim at providing answers or even tools to answering moral dilemmas, they describe the moral person. The main issues in virtue ethics is identifying what the relevant virtues are and how much of each is good, recognizing that virtues in excess may become vices (Hursthouse, 1999). Naturally, actions of people are considered morally correct if they embody the virtues required by the particular school; i.e. a moral dilemma can be answered by considering what a virtuous person would do in such situation. Therefore, there are no forbidden activities, so there is no tenet, for instance, not to kill, the dilemma is whether a virtuous person would kill.

After flourishing in Antiquity, virtue ethics was superseded by other schools, but it did have a revival in the 20-21 centuries, the leading figure being Alasdair MacIntyre (1998) and Elizabeth Anscombe (1958). MacIntyre (2007) also emphasized that we may need new models of virtue ethics for the modern world. Throughout history, and particularly during modernity, with its emphasis on scientific thinking and reductionism, attempts were made to derive all virtues from a single one, so can we say that all virtues are just consequences of e.g. courage or patience or, at least, construct a list of priority, i.e. virtues should be ordered by importance (or other organizing principle). Some recent philosophers, including Julia Annas (2011) and Anscombe (1958) argue that the virtues are all interconnected forming a complex system although the idea can be traced back to Aristotle (Gottlieb, 1984). They also discuss limitations of how actions of virtuous people may occasionally be wrong, which means that being virtuous is not a guarantee of right action. There are many other smaller schools of virtue ethics that we do not cover here. All schools of virtue ethics agree on two principles: they all attribute agency to humans, so that they can exercise their free will and choose to do the right thing, and they all consider humans rational and thus exercising rationality is part of being virtuous.

While areatic schools of ethics focus on what a virtuous person is like, deontic schools of ethics focus on what constitutes a morally right action. Deontology and consequentialism are both deontic (see Figure 1).

Deontology, in essence, offers sets of rules that we must follow to adhere to a high moral standard. Although there were a few earlier versions, the first comprehensive treatment of deontology comes from Immanuel Kant (1785, 1797). The central concept of Kant's deontology is the categorical imperative, according to which we should do what we would be happy to become a universal law. Another very important concept of Kant's deontology is the good will, meaning that if one has the right intentions, but things turn out badly due to things beyond one's control, that is acceptable. This served as basis for the development of a minority branch of deontology, which emphasizes intentions in contrast to consequences. While in these schools consequences do not matter, in most schools they do, only not as much as obligations and rights. A deontological model can incorporate one or both of two perspectives: agency and patiency. Agency links to obligations, i.e. the rules are based on what one is supposed to do. Patiency links to rights, i.e. the rules are based on how people can expect to be treated.

While for Kant rules are universally and infinitely valid, many approaches to deontology are more situational; sometimes it is OK to deviate from the rule for the sake of a better outcome. For instance, it is wrong to lie, but sometimes it may be the kind thing to do, and in some extreme situations (e.g. a monster asking you where your children are because it wants to eat them) it may be the only right thing to do. Of course, exceptions to rules can be formulated as rules, but only so far as we are able to foresee idiosyncratic situations. At first sight, deontology seems an excellent candidate for being considered in AI, as it is rule-based, and rules are easy to program. However, exceptions tend to be problematic in programming as much as in philosophy.

Consequence ethics, like deontology, is concerned with the right action, but not in terms of what we ought to do but in terms of the consequences of our actions. Consequences are typically expressed in the form of some utility or happiness at an individual or social level. Variants of consequentialism include hedonism, egoism, act consequentialism, and various forms of utilitarianism (e.g. Mill, 1861; Williams, 1993). There are two initial issues surrounding consequentialism: (1) Normally, we cannot know the consequences of our actions for sure at the time of taking the action. (2) Desirable consequences are often formulated as the best outcome for the greatest number of people, and thus the rights of minorities may be overlooked. Nevertheless, consequentialism seems to be, at least on a cursory look, to be well aligned with the idea of machine learning

(ML) in artificial neural networks (ANN). Once we take further in-depth considerations, however, we find that there are other aspects of moral decisions that matter besides the consequences, even if we could be sure of those consequences; examples include the trolley problem and similar dilemmas (Foot, 1978; Williams, 1985).

There were a few individual approaches to ethics by leading thinkers of Enlightenment. These thinkers, including Georg Wilhelm Friedrich Hegel, Baruch Spinoza, David Hume, and Adam Smith emphasized the interconnectedness of people, leading to the view that other people are key to one's moral decisions. These approaches can be seen as a move from a fully individualist treatment of ethics towards a social perspective; for example, in a social variant of consequentialism we can consider what is good for society rather than the individual. An implication of these social approaches is the previously mentioned role of patiency in addition to agency in ethical considerations. A more contemporary example of a social view of ethics is Emmanuel Levinas (1961, 1991).

Besides these main normative schools, there are lesser-known ones, many of which can be, in some ways, considered "more human" than the previously introduced major schools. *Pragmatist ethics*, although not very prominently featured in the overall pragmatist philosophy, does not believe in the possibility of a single model that governs all moral decisions; ethics models can be useful, but different ones may be appropriate for different occasions (see e.g. James, 1899). According to *intuitionist ethics* we can know moral principles intuitively, as they are self-evident. These principles, typically duties (which makes intuitionism a form of deontology), are identified individually and therefore intuitionist ethics may be applicable in a greater number of situations than other approaches. For *contractualist ethics*, which is a form of deontology, justice makes an action right – although this is typically formulated in negative terms, i.e. avoiding wrongdoing and injustice. Finally, *feminist ethics* questions many assumptions of normative ethics models, which have all been conceptualized from a dominantly male, often patriarchal, perspective. This leads to rejecting the possibility of any absolute model in ethics (Gilligan, 2014), which means that both pragmatist and feminist ethics become pluralist and contextual.

We do not discuss the variants of ethical approaches in the Frankfurt School, as they are very fragmented and highly blended with political philosophy. What they agree about is similar to the starting point of feminist ethics, i.e. they all suggest that Western ethics is imbued with the exploitative values of the capitalist hegemony, making social justice impossible.

2.1. Moral Compasses

Moral psychology is about what people actually do rather than what they are supposed to do, therefore it can also be labelled as descriptive ethics in contrast with the normative schools (see Figure 1). The term moral psychology can be traced back to Anscombe (1958), who observed that the thinking in ethics should take into consideration what we learned studying psychology.

In one of the cornerstone works of moral psychology James Rest (1986) distinguishes four stages of moral decision making: moral sensitivity, moral judgment, moral motivation, and moral courage. A thorough treatment of moral psychology would require covering a significant amount of psychology literature, both conceptual and experimental, and making connection to the previously outlined normative approaches to ethics. Therefore, here we focus solely on one particularly important issue. Using Rest's stages, psychologists of ethics have established that people make moral judgments more or less exclusively using their intuition, and they only refer to the ethical models that they are familiar with when they need to justify their moral judgments to themselves or to others (cf Haidt, 2001). This is a very strong claim and moral psychology is largely in agreement about this point (Sonenshein, 2007). This is what we tried to capture with the notion of the *moral compass*.

However, we also must note that this does not mean that all ethicists, let alone scholars in the domain of AI ethics, subscribe to the dominant role of intuition in moral judgments; many emphasize deliberation with or without the use of normative models. There are also rare studies that try to synthesize the normative and descriptive approaches to ethics (Treviño, 1986). In our view, humans often arrive at moral judgments intuitively, just as in the case of any decision making, but this can also happen through sequential reasoning (Dörfler & Stierand, 2017). However, if consulting the models we are familiar with does not help to arrive at a moral judgment, we may revert to the use of intuiting as time may be pressing. A case can also be made that the *deliberation before action* is also only employed to justify the *intuitive judgment* already made. The debate is still ongoing, and we will not resolve it in this paper; for us it is important that intuitive judgments exist and both experimental and observational studies find significant use of intuition, particularly at a high level of mastery (Chase & Simon, 1973a; e.g. Chase & Simon, 1973b; Dörfler et al., 2009; Dreyfus, 2004; Dreyfus & Dreyfus, 1986; Gobet & Simon, 1996a, 1996b, 2000; e.g. Kreisler & Dreyfus, 2005). It is reasonable then to assume that moral judgments, like judgments more generally can be intuitive.

3. Ethics and AI

Ethical considerations in computer science are not new, and they get amplified in the world of AI. Norbert Wiener's (1960, p. 1358) formulation is still valid, perhaps more than ever:

"If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it."

The literature on ethics in the scope of computers and wider digitalization, sometimes also referred to as digital ethics, largely applies normative models of ethics to the scope of the digital world (Anderson, 2011; Anderson & Anderson, 2011; Brey, 2000; Flanagan et al., 2008; Friedman et al., 2013; Moor, 1985; Vallor, 2016; van Wynsberghe, 2013). For example, Shannon Vallor (2016) seeks to adapt Aristotelian virtue ethics for the digital future.

The most important issue of AI ethics, as a scholarly discipline, is that it is almost completely conceptual. There are great explorations of applying a variety of normative ethics models in different digital/computerized/AI environments and theorizing or problematizing what the consequences would be – typically not leading to happy conclusions with the logical outcome that we may need a new normative model (Gunkel, 2017). Important problem areas in AI ethics, with reference to normative ethics, are agency, the roles emotions may play, levels of relativism, rationality, more specifically that there are different kinds of rationalities, the use of intuiting, as well as the relationship between a moral decision and action. For instance, rules-based ethics seems to be particularly suitable for computers, but whose rules to accept?

At the same time, AI vendors struggle with the ethical aspects of their products, and they keep looking at AI scholars and philosophers for help that they fail to provide. They look at the potential users of their products before designing a new product as well as after, and the opinions that they receive contradict each other and cannot be programmed. This problem, observed in the reality of AI vendors, is our starting point.

The most popular form of AI today, the deep neural networks (DNN) capable of deep learning (DL) cannot help. Much of the AI success today is ascribed to this form of AI, however, those successes must be looked at in context: the type of problems DL was applied to. In principle, DNN is simply an ANN with more than one

hidden layer, and DL is a really efficient form of ML in a DNN, but the principles are not substantially different: ML needs a large number of learning examples and then it replicates the statistical frequencies of the outcomes with reference to the input variables (LeCun et al., 2015; Marcus, 2018; Schmidhuber, 2015). AlphaGo (Silver et al., 2016) needed some 300 billion games to get trained and deliver the extraordinary performance of beating Lee Sedol (Hassabis, 2017). So, if we even had a database of moral decisions, what number of learning examples would be needed for DL to produce the statistical frequencies? What variables would need to be considered? How many types of moral decisions are there? The list of questions could continue, each and every one of them would be sufficient to conclude that this is not the way to successfully deal with AI ethics. Furthermore, there is evidence that if the training data is biased, the ANN will amplify these biases.

Essentially, we need to understand that the problem of AI ethics is not an implementation problem. It is not about having the right conceptual construct that we need to operationalize, we are struggling with fundamental problems of ethics. Therefore, we need to step back and look into making progress in the field of ethics in which both humans and AI are present. So, what can be done? In order to figure this out, we look into three aspects of how AI is different from the human mind.

4. Methodological Considerations

We take a phenomenological approach in this conceptual study, framed within the broadly considered paradigm of Critical Interpretivism (Dörfler, 2023b). This means that our approach is moderate subjectivist, and we practice bracketing through transpersonal reflexivity to arrive at insights (Dörfler & Stierand, 2021). We do not adopt a theoretical lens, as any lens limits what can be seen, instead we adopt the approach known as phenomenon-driven theorizing, which allows us to approach the phenomenon at hand with an open mind and allow theorizing to take us in various directions (Fisher et al., 2021; Langley, 2021; Ployhart & Bartunek, 2019). The particular type of theorizing we undertake is called problematizing, as the purpose of it is not to provide a solution but to arrive at an improved understanding of what the problem is. To problematize AI ethics, we make use of everyday well-known phenomena, as it is often done in Gestalt psychology (Köhler, 1959; Rock & Palmer, 1990), and attempt to explain these employing abductive reasoning (Sætre & Van de Ven, 2022).

5. Indwelling

Sensing, i.e. sensory input is an indispensable aspect of all knowing and is necessarily employed on par with intellect (Bas et al., 2022). Michael Polányi (1966b, p. 15), the renowned philosopher of knowledge, argues that the body is the ultimate instrument of all external knowledge. Antonio Strati (2007) wonders why the important role of the body is neglected and often ignored although it is the body that enables both intellectual reasoning and sensory-based knowledge. Sensing is not a unitary construct. Based on Burton (2009, p. 37) Dörfler and Bas (2020) consider, besides perception based on the five primary senses, also visceral sensations (e.g. hunger), affective sensations (e.g. love), as well as mental sensations (e.g. pride). With this expanded view of sensing, we can easily conclude that everything we know we come to know through sensing (Bas et al., 2022; de Rond et al., 2019; Strati, 2007). Furthermore, based on Polányi, we introduce the notion of indwelling, through the use of which the idea of sensing can be extended to abstract domains, such as mathematics or astrophysics, or microbiology: these are all abstract in the sense that we cannot get in touch with the subject of inquiry through our body, but the phenomenon is essentially the same. We must emphasize that we do not argue for the empirical over the rational, we suggest considering indwelling in addition to, rather than instead of, reasoning.

Why is this so important? Of course, people sense and there are various mechanical, electronic, etc. sensors that we can connect to computers. Surely a camera with the right set of calculations is more reliable than a human eye... However, two harsh critiques of AI, Hubert Dreyfus (1992) and John Searle (1994) both consider the computers' lack of sensory capacity in producing knowledge as one of the main reasons that computers cannot think and that the "strong AI" paradigm is impossible. A full treatment of this issue would entail exploring the issue of primary and secondary qualities, originally introduced by René Descartes (1637) and then elaborated by John Locke (1690) and later George Berkeley (1878), and deriving the notion of "felt sense" from these (Dörfler, 2023b). Therefore, we do not intend to engage in a generic debate about the possibilities of human-level (or nearly human-level) AI; we simply acknowledge the significance of sensory grounding in ethics in the light of the four phases of moral decision-making identified in moral psychology. As before, we emphasize that we are not arguing that sensing should replace reasoning but that sensing and reasoning are both essential.

6. Understanding

There is no complete agreement in cognitive psychology or the philosophy of mind regarding the precise definition of the concept of understanding. Russell Ackoff (1989) locates it between knowledge and wisdom. We know that tacit knowledge plays a crucial role in understanding, as Polányi (1966a, p. 7) suggests:

“While tacit knowledge can be possessed by itself, explicit knowledge must rely on being tacitly understood and applied. Hence all knowledge is either tacit or rooted in tacit knowledge. A wholly explicit knowledge is unthinkable.”

The significance of understanding in ethics is perhaps obvious: if we are to make moral decisions, we need to be able to understand that our actions have consequences (consequentialism), even if we do not exactly know what the consequences are, as need to understand the rules that we are meant to follow (deontology), we need to understand the implications of specific virtues on our actions (virtue ethics).

Debates on whether AI has or can have understanding are as old as AI. In 1957 Herbert Simon predicted four things AI was supposed to achieve within ten years (Simon & Newell, 1958, pp. 7-8). The only one that has been achieved is that a computer has beaten the best chess player in the world, but that only happened in 1997. Simon has also asserted:

“I believe that in our time computers will be able to perform any cognitive task that a person can perform. I believe that computers already can read, think, learn, create...” (Simon, 1977, p. 6)

Some recent events, however, may rejuvenate these discussions. A chess robot broke a seven-year-old-boy’s finger (Henley, 2022). The Bing AI chatbot called a CNN reporter “rude and disrespectful”, presumably for asking too many questions (Kelly, 2023), and declared love to a NYT journalist and tried to convince him that he did not love his wife, but the chatbot (Roose, 2023). There are also numerous examples of factual and logical mistakes made by ChatGPT-4. However, there was a particularly instructive story that happened in February 2023. Kellin Pelrine, an American amateur Go player (ranked one level below the top in amateurs) beat the top Go computer 14 out of 15 games (Waters, 2023). What makes this point significant, is that all 14 times it was the same trick. If the computer had any level of understanding of the game, it would have identified the same trap being set the second time, let alone being tricked 14 times the same way. However, these examples only showcase that AI does not possess understanding right now, but not that it cannot.

Of course, there are many examples of generative AI (and other forms of AI) delivering incredible performance, the question is whether AI can understand and if not now, can AI ever understand (Chomsky et al., 2023). We believe that the examples conclusively prove that AI does not understand. We also believe that AI is not designed to think, but to mimic some of the outcomes of thinking. Clearly, there are other opinions, in the end it all boils down to whether we accept the computational model of the mind. The view of understanding has a significant impact on the view of AI ethics, as moral judgments presume understanding.

7. Doubting

Finally, the third pillar of our argument is the capacity to doubt. When we paraphrase René Descartes, we accentuate that he was not a Sceptic (one of the four Hellenistic philosophical traditions, of which David Hume was a late follower). Descartes was trying to combat the sceptic dictum that we can doubt everything by attempting to find solid ground in those things that we can be really sure about. To this end, he adopted the Sceptic armament and applied it to everything that he could think of, in a systematic doubt. In doing so, he behaved like a Sceptic, demonstrating how we can doubt anything that we think we know. We cannot be sure that it is an object in front of us, we cannot even be sure of our own bodies, it can all be an illusion, an evil demon’s doing who hijacked our minds. And then, in a master-stroke, he turns the argument upside down, and concludes that if one can doubt anything and everything, then there is one thing that one can be sure of, and it is that there *is* something that can doubt. As doubting is a form of thinking, Descartes (1637, p. 27) formulates his famous dictum: “*Cogito ergo sum*” (I think and therefore I am). However, it is more precise if we limit the term to doubting, and thus ‘*Dubito ergo sum*’ (I doubt and therefore I am). What we are suggesting here, is that being moral entails the capacity to doubt, we need to be able to doubt our actions, to doubt ourselves (de Crescenzo, 1992; e.g. Spiegelberg, 1947).

So, what does it mean to be able to doubt, particularly in the context of ethics? It entails sensing our decision situation and understanding it, trying to do the right thing but not being able to figure out what our actions may lead to, reflecting and not being sure even of our motivations. Just think of Hamlet’s painful struggle whether he should avenge his father. Doubt incorporates some of the most complex issues of the human mind, including sensing and understanding, and it may well be indispensable for our moral decisions, for our moral development, perhaps the central component of the moral mind. There is no consensus about this point, the significance of doubt is our own observation

in the realm of AI ethics. Doubt also has an interesting implication regarding certainty: there is a significant body of literature on uncertainty in entrepreneurship, strategy, and decision making, since Knight (1921, 1923) suggested that our default condition is uncertainty, in which alternatives and their respective probabilities are not known. To cope with uncertainty, we make social contracts, perhaps we can think of doubt in a similar vein. Doubt scarcely appears in the AI literature (see Shklovski & Némethy, 2023 for a rare example). Importantly for AI ethics, if doubt seem to be essential for our moral judgments, what are the implications of doubt-less AI? Importantly, while we developed our argument from a phenomenological perspective, the concept of doubting can be significant for AI in any philosophical position.

8. Final commentary

In several ways, AI ethics is a weird concept. It can cover moral considerations of making AI, ethical aspects of using AI, potential consequences of the tasks we assign to AI, and so forth (Asaro, 2006; Floridi & Sanders, 2002). When we build AI, if we hand over some of our decisions to it, we need to put in something that takes care of those aspects that constitute the moral dimensions of our decisions. How can we do that? What should it be? We know, for instance, that if trained on biased data, AI may amplify those biases.

We will not pretend to have figured out how to build a moral engine for AI. However, we now perhaps understand a little better what needs to be considered for such an attempt. This is incredibly timely, it was as we were writing the first version of this paper that we found out that Microsoft has sacked all its AI ethics team just as they were getting ChatGPT into Bing and soon possibly into many other products. This means one thing for us: ethical questions of AI are difficult, complex, ignite heated debates – and it is paramount that we get them right.

It looks like there is no easy way that would let us ‘program’ ethics into AI, or let it learn it through ML/DL. The reason is that we do not have a proverbial perfect moral entity whose moral characteristics or actions we could use as starting points. Even if we could find such an entity, we would struggle to identify a sensible number of learning examples – and we could not even start figuring this out, as we have no idea how many types of moral decisions exist. There is also no large model that could be used by generative AI.

What does this leave us with? Well, there are a few things we can pin down: (1) The main point of moral psychology was that we make at least some of our moral judgments intuitively and only use ethics models to explain them. (2) Our moral judgments are rooted in

indwelling (typically sensory perception) and we need to understand the decision situation as well as the possible consequences of our actions. This does not mean that we can know the consequences, but we can think up scenarios. (3) The capacity to doubt seems an indispensable part of moral decisions. None of this, however, suggests to us that AI ethics is a futile area.

What we are trying to figure out is what AI can help us with in terms of moral decisions and to understand the way forward. There are two immediate things that we believe AI can do for us right now: (1) AI can provide us with useful input for our process of self-doubt (ex ante or ex post), as we are deliberating our moral decisions by identifying potentially relevant patterns in available ethics models. This can help in two ways, it can reduce the struggle of self-doubt and can help us explain our moral judgments. (2) AI can scan the context for emerging information and patterns, feeding this back to us so that we can course-correct quickly. Humans rely on their ‘felt sense’, like babies calling for their parents when they need a change of diapers: they feel uncomfortable. As AI lacks ‘felt sense’, we need to figure out how to provide external pointers when AI needs to involve a human in the process.

Considering the previous discussion we now make a leap and suggest something that does not trivially follow from what has been said. AI is an amplifier. It does not make us smarter, it amplifies what we have, and if we are stupid, it will amplify that as well (Dörfler, 2022). We have seen e.g. how AI can amplify biases. However, the leap is the following: we suggest that we do not actually have AI ethics problems. This is the reason that we went all the way back and took a journey in time for 2.5 millennia. We do not have *AI ethics problems*, we have *ethics problems*. AI amplifies them.

9. References

- Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, 16(1), 3-9.
- Anderson, S. L. (2011). Philosophical Concerns with Machine Ethics. In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 162-167). Cambridge University Press.
<https://doi.org/10.1017/CBO9780511978036.014>
- Anderson, S. L., & Anderson, M. (2011). A Prima Facie Duty Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a Dialogue with Ethicists. In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 476-492). Cambridge University Press.
<https://doi.org/10.1017/CBO9780511978036.032>
- Annas, J. (2011). *Intelligent Virtue*. Oxford University Press.
- Anscombe, G. E. M. (1958). Modern Moral Philosophy. *Philosophy*, 33(124), 1-19.
<https://doi.org/10.1017/S0031819100037943>

- Aristotle. (cca 330 BCE/1980). *The Nicomachean Ethics* (W. D. Ross, Trans.; W. D. Ross, J. O. Urmson, & J. L. Ackrill, Eds.). Oxford University Press.
- Asaro, P. M. (2006). What Should We Want From a Robot Ethic? *International Review of Information Ethics*, 6(12), 9-16.
- Balasubramanian, N., Ye, Y., & Xu, M. (2020). Substituting Human Decision-Making with Machine Learning: Implications for Organizational Learning. *Academy of Management Review*, 0(ja), null. <https://doi.org/10.5465/amr.2019.0470>
- Bas, A., Sinclair, M., & Dörfler, V. (2022). Sensing: The Elephant in the Room of Management Learning. *Management Learning*. <https://doi.org/10.1177/13505076221077226>
- Berkeley, G. (1878/2008). *A Treatise Concerning the Principles of Human Knowledge* (C. Porterfield Krauth, Trans.). J.B. Lippincott & Company.
- Brey, P. (2000). Disclosive computer ethics. *SIGCAS Comput. Soc.*, 30(4), 10-16. <https://doi.org/10.1145/572260.572264>
- Burton, R. A. (2009). *On being certain: Believing you are right even when you're not*. St. Martin's Press.
- Chase, W. G., & Simon, H. A. (1973a). The Mind's Eye in Chess. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 215-281). Academic Press.
- Chase, W. G., & Simon, H. A. (1973b). Perception in Chess. *Cognitive Psychology*, 4(1), 55-81. [https://doi.org/10.1016/0010-0285\(73\)90004-2](https://doi.org/10.1016/0010-0285(73)90004-2)
- Chomsky, N., Roberts, I., & Watumull, J. (2023, 8th March 2003). Noam Chomsky: The False Promise of ChatGPT. *New York Times*. <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- de Crescenzo, L. (1992). *Il Dubbio [The Doubt]*. Arnoldo Mondadori.
- Csaszar, F. A., & Steinberger, T. (2022). Organizations as Artificial Intelligences: The Use of Artificial Intelligence Analogies in Organization Theory. *Academy of Management Annals*, 16(1), 1-37. <https://doi.org/10.5465/annals.2020.0192>
- Davenport, T. H. (2018). *The AI Advantage: How to Put the Artificial Intelligence Revolution to Work*. MIT Press.
- Davenport, T. H., & Ammanath, B. (2020, 12th August 2020). Redefining AI Leadership in the C-Suite. *Sloan Management Review*. <https://sloanreview.mit.edu/article/redefining-ai-leadership-in-the-c-suite/>
- Davenport, T. H., & Euchner, J. (2023). The Rise of Human-Machine Collaboration. *Research-Technology Management*, 66(1), 11-15. <https://doi.org/10.1080/08956308.2023.2142435>
- Davenport, T. H., & Glaser, J. P. (2022). Factors governing the adoption of artificial intelligence in healthcare providers. *Discover Health Systems*, 1(4). <https://doi.org/10.1007/s44250-022-00004-8>
- Davenport, T. H., & Glover, W. J. (2018). Artificial intelligence and the augmentation of health care decision-making. *NEJM Catalyst*, 4(3). <https://catalyst.nejm.org/ai-technologies-augmentation-healthcare-decisions/>
- Davenport, T. H., & Miller, S. M. (2022). *Working with AI: Real Stories of Human-Machine Collaboration*. MIT Press.
- Davenport, T. H., & Mittal, N. (2022). How Companies Can Prepare for the Coming "AI-first" World. *Strategy & Leadership, ahead-of-print*(ahead-of-print). <https://doi.org/10.1108/SL-11-2022-0107>
- Davenport, T. H., & Mittal, N. (2023). *All-in On AI: How Smart Companies Win Big with Artificial Intelligence*. Harvard Business Review Press.
- Descartes, R. (1637/1986). Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences. In R. Descartes (Ed.), *A Discourse on Method, Meditations and Principles*. Dent.
- Dörfler, V. (2020). Artificial Intelligence. In M. A. Runco & S. R. Pritzker (Eds.), *Encyclopedia of Creativity* (3rd ed., pp. 57-64). Academic Press. <https://doi.org/10.1016/B978-0-12-809324-5.23863-7>
- Dörfler, V. (2022). *What Every CEO Should Know About AI*. Cambridge University Press.
- Dörfler, V. (2023a). Artificial Intelligence. In J. Mattingly (Ed.), *The SAGE Encyclopedia of Theory in Science, Technology, Engineering, and Mathematics* (Vol. 1, pp. 37-41). Sage. <https://doi.org/10.4135/9781071872383.n15>
- Dörfler, V. (2023b). *Critical Interpretivism: The First Outline PHILOS 2023: 3rd Colloquium on Philosophy and Organization Studies*, Chania, Greece
- Dörfler, V., Baracska, Z., & Velencei, J. (2009, 7-11 August 2009). *Knowledge Levels: 3-D Model of the Levels of Expertise AoM 2009: 69th Annual Meeting of the Academy of Management*, Chicago, IL. <https://www.researchgate.net/publication/308339223>
- Dörfler, V., & Bas, A. (2020). Intuition: scientific, non-scientific or unscientific? In M. Sinclair (Ed.), *Handbook of Intuition Research as Practice* (pp. 293-305). Edward Elgar. <https://doi.org/10.4337/9781788979757.00033>
- Dörfler, V., & Stierand, M. (2017). The Underpinnings of Intuition. In J. Liebowitz, J. Paliszkievicz, & J. Gołuchowski (Eds.), *Intuition, Trust, and Analytics* (pp. 3-20). Taylor & Francis. <https://doi.org/10.1201/9781315195551-1>
- Dörfler, V., & Stierand, M. (2021). Bracketing: A Phenomenological Theory Applied Through Transpersonal Reflexivity. *Journal of Organizational Change Management*, 34(4), 778-793. <https://doi.org/10.1108/JOCM-12-2019-0393>
- Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason* (revised ed.). MIT Press.
- Dreyfus, H. L. (2004). *A Phenomenology of Skill Acquisition as the basis for a Merleau-Pontian Nonrepresentationalist Cognitive Science*. University of California, Department of Philosophy. <https://philpapers.org/archive/DREAPO.pdf>
- Dreyfus, H. L., & Dreyfus, S. E. (1986/2000). *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. The Free Press.
- Fisher, G., Mayer, K., & Morris, S. (2021). From the Editors - Phenomenon-Based Theorizing. *Academy of*

- Management Review*, 46(4), 631-639.
<https://doi.org/10.5465/amr.2021.0320>
- Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying Values in Technology: Theory and Practice. In J. van den Hoven & J. Weckert (Eds.), *Information Technology and Moral Philosophy* (pp. 322-353). Cambridge University Press.
<https://doi.org/10.1017/CBO9780511498725.017>
- Floridi, L., & Sanders, J. W. (2002). Mapping the Foundationalist Debate in Computer Ethics. *Ethics and Information Technology*, 4(1), 1-9.
<https://doi.org/10.1023/A:1015209807065>
- Foot, P. (1978). *Virtues and Vices and Other Essays in Moral Philosophy*. University of California Press.
- Friedman, B., Kahn, P. H., Borning, A., & Hultgren, A. (2013). Value Sensitive Design and Information Systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), *Early engagement and new technologies: Opening up the laboratory* (pp. 55-95). Springer Netherlands. https://doi.org/10.1007/978-94-007-7844-3_4
- Gilligan, C. (2014). Moral Injury and the Ethic of Care: Reframing the Conversation about Differences. *Journal of Social Philosophy*, 45(1), 89-106.
<https://doi.org/10.1111/josp.12050>
- Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14(2), 627-660.
<https://doi.org/10.5465/annals.2018.0057>
- Gobet, F., & Simon, H. A. (1996a). Recall of Random and Distorted Chess Positions: Implications for the Theory of Expertise. *Memory & Cognition*, 24(4), 493-503.
<https://doi.org/10.3758/BF03200937>
- Gobet, F., & Simon, H. A. (1996b). Templates in Chess Memory: Mechanism for Re-calling Several Boards. *Cognitive Psychology*, 31(1), 1-40.
<https://doi.org/10.1006/cogp.1996.0011>
- Gobet, F., & Simon, H. A. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science*, 24(4), 651-682. [https://doi.org/10.1016/S0364-0213\(00\)00031-8](https://doi.org/10.1016/S0364-0213(00)00031-8)
- Göndöcs, D., & Dörfler, V. (2022, 7-9 July 2022). *AI-enabled Organizational Learning Strategy EGOS 2022: 38th Colloquium of the European Group for Organization Studies*, Vienna, Austria.
<https://www.researchgate.net/publication/363049580>
- Göndöcs, D., & Dörfler, V. (2023, 7-14 February 2023). *AI in Medical Diagnosis: AI Prediction vs Human Judgement AAAI 2023: 37th AAAI (Association for the Advancement of Artificial Intelligence) Conference on Artificial Intelligence*, Washington, DC.
<https://www.researchgate.net/publication/368779553>
- Gottlieb, P. (1984). Aristotle on Dividing the Soul and Uniting the Virtues. *Phronesis*, 39(3), 275-290.
<https://doi.org/10.1163/156852894321052081>
- Grodal, S., Krabbe, A. D., & Chan-Zninog, M. (2023). The Evolution of Technology. *Academy of Management Annals*, 17(1), 141-180.
<https://doi.org/10.5465/annals.2021.0086>
- Gunkel, D. J. (2017). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814-834.
<https://doi.org/10.1037/0033-295X.108.4.814>
- Hassabis, D. (2017). *The Future of AI* New Scientist Live 2017, London, UK.
- Henley, J. (2022, 24th July). Chess robot grabs and breaks finger of seven-year-old opponent. *The Guardian*.
<https://www.theguardian.com/sport/2022/jul/24/chess-robot-grabs-and-breaks-finger-of-seven-year-old-opponent-moscow>
- Hursthouse, R. (1999). *On Virtue Ethics*. Oxford University Press.
- James, W. (1899/2001). On a Certain Blindness in Human Beings. In W. James (Ed.), *Talks to Teachers on Psychology: And to Students on Some of Life's Ideals*. Dover.
- Kant, I. (1785/1998). *Groundwork of the Metaphysics of Morals* (M. Gregor, Ed.). Cambridge University Press.
- Kant, I. (1797/2017). *The Metaphysics of Morals* (M. Gregor, Trans.; L. Denis, Ed. revised ed.). Cambridge University Press.
- Kelly, S. M. (2023, 16th February). The Dark Side of Bing's New AI Chatbot. *CNN*.
<https://www.cnn.com/2023/02/16/tech/bing-dark-side>
- Knight, F. H. (1921). *Risk, Uncertainty and Profit*. Houghton Mifflin.
- Knight, F. H. (1923). Business Management: Science or Art? *Journal of Business*, 2(4th March), 5-24.
- Köhler, W. (1959). Gestalt Psychology Today: Address of the President at the sixty-seventh Annual Convention of the American Psychological Association, Cincinnati, Ohio, September 6, 1959. *American Psychologist*, 14(12), 727-734. <https://doi.org/10.1037/h0042492>
- Kreisler, H., & Dreyfus, H. L. (2005). *Meaning, Relevance, and the Limits of Technology: Interview with Hubert Dreyfus* (Conversations with History, Issue).
<http://globetrotter.berkeley.edu/people5/Dreyfus/>
- von Krogh, G. (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, 4(4), 404-409.
<https://doi.org/10.5465/amd.2018.0084>
- Langley, A. (2021). What Is "This" a Case of? Generative Theorizing for Disruptive Times. *Journal of Management Inquiry*, 30(3), 251-258.
<https://doi.org/10.1177/10564926211016545>
- Leavitt, K., Schabram, K., Hariharan, P., & Barnes, C. M. (2021). The Machine Hums! Addressing Ontological and Normative Concerns Regarding Machine Learning Applications in Organizational Scholarship. *Academy of Management Review*, 0(ja), null.
<https://doi.org/10.5465/amr.2021.0166>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
<https://doi.org/10.1038/nature14539>
- Levinas, E. (1961/1991). *Totality and Infinity: An Essay on Exteriority*. Kluwer Academic Publishers.
- Levinas, E. (1991/2017). *Entre Nous* (B. Harshav & M. B. Smith, Trans.). Bloomsbury.

- Lindebaum, D., & Ashraf, M. (2021). The Ghost in the Machine, or the Ghost in Organizational Theory? A Complementary View on the Use of Machine Learning. *Academy of Management Review*, 0(ja), null. <https://doi.org/10.5465/amr.2021.0036>
- Locke, J. (1690/1959). *An Essay Concerning Human Understanding* (Vol. 1). Dover Publications.
- MacIntyre, A. (1998). *A Short History of Ethics: A History of Moral Philosophy from the Homeric Age to the Twentieth Century* (2nd ed.). Routledge.
- MacIntyre, A. (2007). *After Virtue: A Study in Moral Theory* (3rd ed.). University of Notre Dame Press.
- Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.
- Mill, J. S. (1861). Utilitarianism. In G. Williams (Ed.), *Utilitarianism, On Liberty, Considerations on Representative Government, Remarks on Bentham's Philosophy* (pp. 1-67). Everyman.
- Moor, J. H. (1985). What Is Computer Ethics? *Metaphilosophy*, 16(4), 266-275. <https://doi.org/10.1111/j.1467-9973.1985.tb00173.x>
- Oliver, N., Calvard, T., & Potočník, K. (2017). Cognition, Technology, and Organizational Limits: Lessons from the Air France 447 Disaster. *Organization Science*, 28(4), 729-743. <https://doi.org/10.1287/orsc.2017.1138>
- Pachidi, S., Berends, H., Faraj, S., & Huysman, M. (2021). Make Way for the Algorithms: Symbolic Actions and Change in a Regime of Knowing. *Organization Science*, 32(1), 18-41. <https://doi.org/10.1287/orsc.2020.1377>
- Ployhart, R. E., & Bartunek, J. M. (2019). Editors' Comments: There Is Nothing So Theoretical As Good Practice – A Call for Phenomenal Theory. *Academy of Management Review*, 44(3), 493-497. <https://doi.org/10.5465/amr.2019.0087>
- Polányi, M. (1966a). The Logic of Tacit Inference. *Philosophy*, 41(155), 1-18. <https://doi.org/10.1017/S0031819100066110>
- Polányi, M. (1966b/1983). *The Tacit Dimension*. Peter Smith.
- Raisch, S., & Krakowski, S. (2021). Artificial Intelligence and Management: The Automation-Augmentation Paradox. *Academy of Management Review*, 46(1), 192–210. <https://doi.org/10.5465/amr.2018.0072>
- Rest, J. R. (1986). *Moral Development: Advances in Research and Theory*. University of Minnesota Press.
- Rock, I., & Palmer, S. (1990). The Legacy of Gestalt Psychology. *Scientific American*, 263(6), 84-91. <http://www.jstor.org/stable/24997014>
- de Rond, M., Holeman, I., & Howard-Grenville, J. (2019). Sensemaking from the Body: An Enactive Ethnography of Rowing the Amazon. *Academy of Management Journal*, 62(6), 1961-1988. <https://doi.org/10.5465/amj.2017.1417>
- Roose, K. (2023). A Conversation With Bing's Chatbot Left Me Deeply Unsettled. *New York Times*. <https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html?smid=url-share>
- Sætre, A. S., & Van de Ven, A. H. (2022). Abductive Theorizing Is More than Idea Generation: Disciplined Imagination and a Prepared Mind [Dialogue]. *Academy of Management Review*, 0(ja), null. <https://doi.org/10.5465/amr.2021.0317>
- Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks*, 61, 85-117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Searle, J. R. (1994/2002). *The Rediscovery of the Mind* (paperback ed.). MIT Press.
- Shklovski, I., & Némethy, C. (2023). Nodes of certainty and spaces for doubt in AI ethics for engineers. *Information, Communication & Society*, 26(1), 37-53. <https://doi.org/10.1080/1369118X.2021.2014547>
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search [Article]. *Nature*, 529, 484. <https://doi.org/10.1038/nature16961>
- Simon, H. A. (1977). *The New Science of Management Decision* (3rd ed.). Prentice-Hall.
- Simon, H. A., & Newell, A. (1958). Heuristic Problem Solving: The Next Advance in Operations Research. *Operations Research*, 6(1), 1-10. <https://doi.org/10.1287/opre.6.1.1>
- Sonenshein, S. (2007). The Role of Construction, Intuition, and Justification in Responding to Ethical Issues at Work: The Sensemaking-Intuition Model. *Academy of Management Review*, 23(4), 1022-1040. <https://doi.org/10.5465/AMR.2007.26585677>
- Spiegelberg, H. (1947). Indubitables in Ethics: A Cartesian Meditation. *Ethics*, 58(1), 35-50. <https://doi.org/10.1086/290589>
- Strati, A. (2007). Sensible knowledge and practice-based learning. *Management Learning*, 38(1), 61-77. <https://doi.org/10.1177/1350507607073023>
- Treviño, L. K. (1986). Ethical Decision Making in Organizations: A Person-Situation Interactionist Model. *Academy of Management Review*, 11(3), 601-617. <https://doi.org/10.5465/amr.1986.4306235>
- Tschang, F. T., & Almirall, E. M. (2021). Artificial Intelligence as Augmenting Automation: Implications for Employment. *Academy of Management Perspectives*, 35(4), 642-659. <https://doi.org/10.5465/amp.2019.0062>
- Vallor, S. (2016). *Technology and the virtues: a philosophical guide to a future worth wanting*. Oxford University Press.
- Waters, R. (2023, 17th February). Man beats machine at Go in human victory over AI. *Financial Times*. <https://www.ft.com/content/175e5314-a7f7-4741-a786-273219f433a1>
- Wiener, N. (1960). Some Moral and Technical Consequences of Automation. *Science*, 131(3410), 1355-1358. <https://doi.org/10.1126/science.131.3410.1355>
- Williams, B. (1985/2006). *Ethics and the Limits of Philosophy*. Routledge.
- Williams, B. (1993). *Morality: An Introduction to Ethics*. Cambridge University Press.
- van Wynsberghe, A. (2013). Designing Robots for Care: Care Centered Value-Sensitive Design. *Science and Engineering Ethics*, 19(2), 407-433. <https://doi.org/10.1007/s11948-011-9343-6>