Embedding Ethics Into Artificial Intelligence: Understanding What Can Be Done, What Can't, and What Is Done

Clayton Peterson
Université du Québec à Trois-Rivières
3351 Bd des Forges, Trois-Rivières (QC), G8Z 4M3
clayton.peterson@uqtr.ca

Abstract

Embedding ethical considerations within the development of AI driven technologies becomes more and more pressing as new technologies are developed. Given the impact of autonomous technologies on individuals and society, it is worth taking the time to assess and manage the ethical aspects and possible consequences of our technological endeavors. While the growing rapidity of autonomous decision processes makes it hard to keep individuals in the decision loops, people are turning their attention to the ways in which ethics could be integrated to machines and algorithms, as well as to the possibility of defining autonomous ethical machines that would be able to solve ethical dilemmas and act ethically (e.g. autonomous vehicles). Notwithstanding theoretical and practical difficulties surrounding the possibility of defining such ethical machines, important elements should be considered when reflecting on the embedding of ethics into AI technologies. The present paper aims to critically analyze the limitations of such endeavors by exposing common misconceptions relating to AI ethics.

Looking Back to the Greeks

A mentor once told me that there is nothing original in contemporary philosophy insofar as everything can be traced back to the Greeks. The Greeks thus seemed a good place to start. One thing he taught was that Plato and Aristotle saw moral relativism (i.e. the idea that objective ethical judgment is impossible) as public enemy #1 at the time. They argued that relativism should be refuted given that it has considerable social and political consequences. Indeed, if ethical judgment is relative to individuals or groups, then there can be no objective rational justification regarding the choices we should make or the actions we should accomplish, either as individuals or as a society. As a consequence, moral relativism has to be refuted in order to allow citizens and policy makers to pursue objective and rational ethical judgment. This position is still topical when looking at the public and academic discourse on artificial intelligence (AI) and technology, with various contradictory opinions being expressed on what should be done. That said, one might argue that public enemy #1 nowadays is rather *ignorance* of what both ethics and AI are, which blinds us to the choices we should make and the actions we should undertake. This paper thus targets ignorance. This might appear as a bold statement, but again think of the Greeks. Socrates, through Plato's youth dialogues, performed *refutations*. By discussing with various authority figures, he questioned them in order to show they did not really know what they were talking about. Although these authority figures took this personally as a wrongdoing, this was really a rhetorical and pedagogical technique: One cannot be open to learning if one thinks one already knows. As such, recognizing our own ignorance is the first step towards knowledge. Hence Socrates' aphorism "All I know is that I know nothing", stating the recognition of his own ignorance and paving the way to his search of wisdom. This is what I intend to do within this paper. Humbly recognizing my own ignorance, I begin at the very beginning and position how the ethics of AI can be conceived in contrast to moral relativism. Then, I discuss how benchmarking cannot be taken as an ethical justification of autonomous vehicles' choices and conclude by discussing how ethical issues can be understood from the perspective of different levels of AI, as well as how consent is a fundamental issue for ethical AI.

Understanding Ethics of AI

Ethics of AI, machine ethics, and ethical AI, can be understood in various complementary ways depending on how one understands both ethics and AI.

Ethics

While ethics can be generally conceived as a systematic and rational evaluation of norms, principles, and values that should constraint and guide our choices, behaviors and actions, there is a distinction to be made between *normative ethics* and *applied ethics*. Normative ethics concentrates on the definition and conceptualization of normative theories meant to regulate ethical choices and behavior. Well-known (and perhaps overemphasized) theories in normative ethics are deontology (e.g. Kant 1785), consequentialism (e.g. Audard 1999; Bentham 1834; Sen and Williams 1982) and virtue ethics (e.g. Anscombe 1958), but the field of normative ethics is composed of various other relevant approaches including care ethics (e.g. Gilligan 1982), ethical minimalism (e.g. Ogien 2007), particularism (e.g. Dancy 2004), or

contractualism (e.g. Rawls 1987), to name only a few. To some extent, the distinction between normative and applied ethics can be seen from the fact that normative ethics studies the justification of principles and values, whereas applied ethics rather studies the justification of concrete actions and choices. Indeed, applied ethics focuses on concrete real-life situations where ethical dilemmas arise and where there are conflicts of norms, values, principles, or obligations. It requires the recognition of ethical (or value) pluralism (see Peterson and Hamrouni 2022; van den Hoven 2010; Weinstock 2017), a point of view emphasizing the existence of multiple complementary (though often incompatible) views on ethical dilemmas. Ethical pluralism can be seen as an (antirealist) answer to moral relativism, which is in itself an answer to the problem of attributing truth values (understood from a realist perspective) to normative statements. Roughly stated, realism posits that truth of statements are established through facts (i.e. correspondence with the empirical world), that knowledge corresponds to having access to these truths, and that these truths can be known. Antirealism denies at least one of these premises. Notwithstanding there is no universally accepted moral theory, there is a conceptual (epistemic) problem that has led some scholars to defend moral relativism (cf. Harman 1978), a view rejecting objective ethical evaluations and defending that ethical evaluations rather depend upon one's perception or conception of ethics, such as cultural values. This epistemic problem can be summarized as follows (cf. Jørgensen 1937): Assuming statements can be true when they correspond to facts, and considering that normative statements do not express how things are but rather are things should be (perhaps in an ideal world), it follows that normative statements cannot be true (at least not as a correspondence with the world we live in).

Plato initially refuted relativism by showing its inconsistency. To summarize, if ethical judgment depends on individuals' conception of ethics, then a same object (e.g. an action) can be thought of as good *and* bad (or not good) at the same time, which is contradictory. However, this argument might not be convincing for a (realist) relativist defending that ethical statements lack truth values (i.e. that they cannot be true or false). If ethical statements can't be true or false, then they can't contradict each other. So even if one sees Plato's argument as sound, and rather thinks that the realist conception of truth and knowledge should be discarded (as I do), it is necessary to provide an argument against relativism that does not rely on truth. One argument can be formulated based on Moore's (1959) open question argument. In a nutshell, assume that, indeed, ethical statements depend on one's or a group's personal or shared values (e.g. cultural values). If we accept that individuals and groups can perform self-criticism, that is, that they can analyze critically their own values and principles in order to evaluate whether they are appropriate or not, then we are denying moral relativism insofar as we are accepting that it is possible to evaluate these values from an external ethical standpoint.[1] There are therefore two options: Either one argues against this possibility, in which case one will not be able to justify social choices regarding what should be done with respect to AI and technologies given that this would imply that ethics does not exist and ethical statements would only be unjustified expressions of personal opinions (think of Aristotle's and Plato's rationale underlying the importance of refuting public enemy #1), or one sees introspection and self-criticism as possible and, therefore, rejects moral relativism (to avoid inconsistency and circularity).

To be clear, Plato's point is that asserting that moral relativism is true is inconsistent insofar as it is a moral theory that states that moral theories can't be true. Relativists are usually missing that point, though. If one rejects the idea that ethical statements are declarative (i.e. if one rejects the idea that ethical statements can be true or false), then the conclusion is not that ethical evaluations depend on individuals or groups: Rather, it follows that ethical evaluations are meaningless and that there is no such thing as *ethics*. Hence, the dilemma is quite simple: Either one really endorses this position, or one does not. Either ethics is, or ethics is not. If ethics is not, then there is no choice regarding what should be done that can objectively be justified. If ethics is, then it is not relative.

Overall, it is not because normative statements (and ethical theories) cannot be true (at least from a realist standpoint) that objective ethical judgment is impossible. Ethical pluralism is a conceptual framework that can properly answer moral relativism by allowing for a plurality of ethically understandable (though incompatible) positions by emphasizing reasonableness rather than truth (Maclure 2020; Pellé and Reber 2016; van den Hoven 2010; Van de Poel and Royakkers 2011; Weinstock 2017). Instead of insisting on the idea that the ethical statements that should be accepted are the ones that are true, it rather understands methodological concerns at the core of an ethical evaluation: The choices that are made need to be justified (more on that topic later) by appealing to principles and values that are objective, in the sense that they are not subjective or idiosyncratic, and universal, in the sense that people not sensible to these values or principles would be characterized as unreasonable (e.g. think of the ideal of a reasonable person in law). Two people can (reasonably) disagree regarding what to do when facing a conflict between two important values such as *security* and *privacy* (e.g. think of the apps developed to track infections during the pandemic), for instance, but we would not characterize these individuals as reasonable if they were to defend that either *security* or *privacy* is worth nothing and should not even be considered. In the end, what matters is not that ethical statements are true or false, but rather that our (ethical) choices and actions are reasonably justified and rational. By rational, we do not only mean consistent (not contradictory), but also instrumental rationality: The choices we make and the actions we accomplish should be explainable by appealing to principles we aim to satisfy or values we aim to reach. While others might disagree with the choices and actions we make, they should minimally be able to *un-*

[1] Assuming that this evaluation is made from another relative standpoint would not solve the problem. For instance, one could consider a normative evaluation of all principles and values, which requires an external standpoint (or risks circularity).

derstand why we acted or chose in such a way. An ethical choice is a choice that can be explained.

Although ethical pluralism can in itself be considered from the perspective of normative ethics (i.e. one can study the justification of principles and values), it can also be used from a practical standpoint to help analyze concrete ethical dilemmas and problems. One interesting aspect of ethical pluralism is that reconciliation of competing normative theories can be accomplished by recognizing the fact that these theories tend to focus on specific aspects of ethical dilemmas. For instance, deontology concentrates on the idea that actions need to intentionally conform to universal norms; consequentialism focuses not on what intentions are but rather on the results of some choices or actions; virtue ethics emphasizes the importance of individuals' character traits; care ethics insists on the well-being of the individuals concerned. All in all, different aspects can be analyzed from an ethical standpoint, including actions, intentions, consequences, norms, risks, as well as AI and technology. As such, different normative theories will tend to focus on different aspects, and ethical pluralism is consistent with the rejection of the thesis that algorithms and machines are value neutral (i.e. that they cannot be the object of an ethical evaluation; cf. Miller 2021). In the end, there is no aspect that is intrinsically superior to others in the ethical evaluation of a situation and, accordingly, there is no normative theory that should always be granted priority.

Artificial Intelligence

As ethics can be understood in different ways, AI can also have different meanings. In the public sphere, for instance, journalists and media often reify or anthropomorphize AI by speaking of *an* artificial intelligence (cf. De Cremer and Kasparov 2022; Ryan 2020). Wooldridge (2021) concisely exposes the issue:

> "the public debate on AI [...] is largely fixated on the grand dream and on alarmist dystopian scenarios that have become a weary trope when reporting on AI ([...] super intelligent AI might go wrong and eliminate humanity). Much of what is published about AI in the popular press is ill-informed or irrelevant (p.3)."

A first distinction to be made is the one between *strong* and *weak* AI (cf. Searle 1980; Wooldridge 2021), that is, between AI that would be self-aware and have a conscious mind and understanding, versus AI that is not. Strong AI is sometimes also referred to as *Artificial Super Intelligence* (cf. Kaplan and Haenlein 2019). Weak AI can be divided into two categories (Kaplan and Haenlein 2019): *Artificial Narrow Intelligence* corresponds to AI defined to apply to specific areas (e.g. facial recognition, translation) but that cannot by itself solve problems in different areas, whereas *Artificial General Intelligence* refers to AI that could be applied to several areas and that could by itself solve problems in other areas (which does not necessarily assume self-awareness or consciousness; cf. Wooldridge 2021). Whether or not strong AI is possible is an issue we will not be addressing within this paper. It suffices to note that this possibility is not trivial, that it has not happened yet, and (although some might disagree) that we are far from achieving it (on the relationship between strong AI and ethics, see Peterson and Hamrouni 2022). As such, we can only but agree with Wooldridge (2021) and others (e.g. De Cremer and Kasparov 2022) in that we should not be reifying or anthropomorphizing AI insofar as this blinds us to more pressing issues arising from what AI actually is, which can be characterized as *narrow* AI (see also Ryan 2020). Indeed, what we do have currently from a technological standpoint is various applications of algorithms (e.g., optimization, learning, predictions) and statistical techniques to different kinds of technologies in specific areas (e.g. facial recognition). According to these distinctions, there are (at least) four ways in which the ethics of AI can be understood:

Strong AI and Normative Ethics One way to understand the ethics of AI is to focus on strong AI from a normative standpoint, e.g. reflecting on notions such as agency, rights, relationships between humans and machines, transhumanism, etc. This reflection is often paired with considerations on the singularity hypothesis (cf. Dehaene, Lau, and Kouider 2021; Torrance 2011; Wolf 2021; Yampolskiy 2013).

Strong AI and Applied Ethics Given that strong AI is speculative and that applied ethics concentrates on concrete cases, this understanding is, for the moment (and perhaps forever), either not applicable or also speculative.

Weak AI and Normative Ethics This approach is characterized by an analysis of the general principles and values that should guide and constraint technological development and usage. This is (and has been) accomplished, for example, through the Montreal Declaration for Responsible AI and the United Nations through their *Recommendations for the Ethics of Artificial Intelligence* (UNESCO 2021).

Weak AI and Applied Ethics This approach insists on analyzing ethical dilemmas and conflicts of norms and values within the context of specific technologies.

Focusing on (weak) AI[2], it should be emphasized that the interpretation of the ethics of AI in terms of *AI and Applied Ethics* suffers in both the public and the academic spheres. Indeed, the focus is usually rather on *AI and Normative Ethics*. In addition to the anthropomorphization and reification of AI in the popular culture and the media, much of the attention in the ethics of AI is given to the analysis and establishment of proper legislation and governance principles, which is in itself a normative activity that is bounded to make an abstract reflection overlooking the specific characteristics of potential cases and problems. In addition, the practical analysis of applied cases is usually seen as a prerogative of companies and industries. As an example, Articles 7 and 8 of Bill C-27 (Minister of Innovation, Science, and Industry 2022), which is meant to regulate development and usage of AI in Canada, states not only that the responsibility to assess whether an AI driven technology falls within the category of 'high-impact systems' is attributed to those who develop this system, but that the ethical evaluation of

[2]From an ethical standpoint, the distinction between *narrow* and *general* AI is not as relevant as the one between *strong* and *weak* AI insofar as both can be analyzed similarly (e.g. design, areas of application, risks, etc.).

their system, including the evaluation of the risks, also does:

> "**Assessment - High-impact system** A person who is responsible for an artificial intelligence system must [...] assess whether it is a high-impact system. **Measures related to risks** A person who is responsible for a high-impact system must [...] establish measures to identify, assess and mitigate the risks of harm or biased output that could result from the use of the system."

In addition, there is a tendency to see the issue of embedding ethics into technologies as a matter of technical competencies and, further, as a problem that will (as if it could) be solved by AI (De Cremer and Kasparov 2022). Embedding ethics into technological development, however, cannot be reduced to a problem of technical competencies, as it cannot be solved by AI (Peterson and Hamrouni 2022). Hence, complementary to the normative analysis of AI, it is also important to analyze the ethical issues surrounding the development of specific technologies from an applied perspective, and this requires proper training in AI ethics.

Choosing Words Knowingly

Properly understanding the ethics of AI requires that some words be given specific attention. Among these words standout two on which I would like to focus now. The first one is *ethics* and its derivatives. When looking at the literature, we stumble upon expressions like *autonomous ethical (moral) agents* (Moor 2006) and *ethical machines* (Anderson and Anderson 2007). These expressions are accompanied by the idea that such machines could be used to unequivocally solve ethical dilemma. Similarly, it is proposed that the behavior and actions of such machines and agents should be accepted because they are *ethical*.

This view, however, is misleading. First, there is no such thing as a unique solution to an ethical dilemma. An ethical dilemma is, by definition, a situation where a conflict between norms, principles, obligations or values requires a sacrifice. If ethical choices are to be explainable through the norms and values one aims to satisfy, then any ethical dilemmas implies *a priori* (at least) two possible 'ethical' choices. Second, to state that something is ethical while something else is not is misleading. What does it mean for a choice or an action to be 'ethical'? Either this means it is 'the' action to be accomplished, in which case one would need to defend a monist view of ethics and argue in favor of 'the' ethical theory (cf. van den Hoven 2010), or one sees this as one possibility among others. What people implicitly do mean when stating that an action or a choice is 'ethical' is that this action or choice conforms to a normative theory, an ethical principle, or some value. What they don't realize is that this does not imply that the action or choice should necessarily be accepted, for there are many theories, principles and values. People can reasonably disagree regarding what to do when facing an ethical dilemma. This is what Rawls (2005, p.441) presented as the "fact of reasonable pluralism". One should thus be aware of what it means to say that a choice, an action, a machine or an algorithm is *ethical*. It is not because it conforms to a normative theory that it should be accepted. Consequentialism, to take a notorious example, suffers from the repugnant conclusion that the needs of the many outweigh those of the few (Parfit 1984). It does not necessarily entail acceptability.

A second word that is usually paired up with ethics and often misconstrued is *justification*, a notion widely spread within the literature on eXplainable AI (XAI). As an illustration of its place within the literature, Moor (2006), for instance, sees full ethical agents as ones that can "reasonably justify [their explicit ethical judgments]", whereas Miller (2019) and Biran and Cotton (2017) see a justification as an explanation of why a decision is good (which is common within the XAI literature; cf. Adadi and Berrada 2018), though Biran and Cotton also refer to an explanation as the description of the rationale behind each step of a decision, which rather corresponds to *propositional justification*. Following Peterson and Broersen (2024), we can distinguish between *ethical justification*, which provides normative reasons supporting why a decision is good (cf. Raz 1999), and *propositional justification* (Turri 2010), which provides reasons explaining why or how a decision occurred.

People tend to think of ethically justified choices as choices they are entitled to make insofar as they are *the* choices to be made. It refers to the implicit idea that *justified choices* conform to an ideal of justice. This is the rationale underlying the idea that actions and choices made by artificial autonomous ethical agents should be accepted. As we saw, however, this understanding is misconstrued. An ethically justified choice is one that can be explained by appealing to ethical norms, values, or principles, that is, norms, values and principles for which it would be unreasonable not to be sensible to. It is ethically justified in the sense that one appeals to normative reasons to support the choice, but this does not imply that the choice can be taken as 'the' choice to be made. As such, an ethically justified choice needs to be understood from the perspective of a propositional justification that in the end appeals to normative reasons.

Let us insist a bit further on this point and expose why it would be a mistake to understand an ethically justified choice as a *good* choice. When facing an ethical dilemma, there is no such thing as a good choice to be made. There are bad answers, to be clear, but at some point all solutions to dilemmas imply a sacrifice. One might argue that the (classical) excluded middle does not hold for good and bad: Something that is not good is not necessarily bad, and vice versa. After all, there are things falling outside of the scope of ethics, and which are per se neither good nor bad. Hence, it would be more accurate to speak of choices that are less bad (rather than better) than others. An ethical dilemma is a non-ideal situation where some value or principle needs to be sacrificed from a pragmatic standpoint if one is to act. Speaking of good choices in that context is debatable. From the perspective of applied ethics, the important point is to know what will be sacrificed and why, that is, the choice needs to be ethically justified following an appropriate evaluation of all the options, principles and values at play.

Benchmarking Autonomous Decisions

Autonomous moral agents are especially present within the context of autonomous vehicles (Faulhaber et al. 2019;

LaCroix 2022; Nascimento et al. 2019; Wang et al. 2020). One way to evaluate whether models perform adequately (i.e., whether they are appropriate or accurate) is by comparing machine learning algorithms' outputs to actual decisions made by individuals (Bjørgen et al. 2018; LaCroix 2022). This way of assessing the (alleged) ethical accuracy of machine learning algorithms is known as *benchmarking*. For instance, through the moral machine experiment conducted at the Massachusetts Institute of Technology between 2016 and 2020, Awad et al. (2018; 2019) studied how people perceived various fictional decisions made by autonomous vehicles in order to establish "socially acceptable principles for machine ethics" (see also Noothigattu et al. 2018). The moral machine experiment gathered people's opinions on examples inspired by variations of Philippa Foot's (1967) seminal trolley problem (i.e. dilemma between who one should save), which were then used to benchmark the appropriateness of decisions made by machine learning algorithms (Faulhaber et al. 2019; Hendrycks et al. 2020; Li et al. 2016; Noothigattu et al. 2018). By doing so, scholars in machine ethics believe they are establishing whether these models are ethical and whether the decisions made were the right ones. The problem, however, is that what people *do* think with respect to the decisions made by fictional autonomous vehicles is insufficient to justify what they *should* think. As LaCroix (2022) argues, scholars in machine ethics are not gathering facts about ethics but are rather gathering sociological facts about people's opinions. Contrary to what scholars in machine ethics are advocating (Faulhaber et al. 2019; Hendrycks et al. 2020; Li et al. 2016; Noothigattu et al. 2018; Sparrow 2004), the type of data that is used as benchmark cannot provide an appropriate basis to justify machine learning algorithms aimed at autonomous ethical decision making (i.e. it cannot provide an ethical justification; cf. Etienne 2022; Peterson and Broersen 2024).

Levels of AI

It is insightful to look at how AI can be characterized in order to better understand how ethical AI can be achieved. AI can be seen to perform at different levels. The American Medical Association (AMA; see the CPT Editorial Panel 2021), for example, distinguishes between three broad categories, including assistive AI (detection without analysis), augmentative AI (provides an analysis to produce a clinically meaningful output) and autonomous AI (interpretation of data accomplished by the machine), the latter category further divided into three subcategories, namely level I (draws conclusion and diagnosis but requires physician to implement), level II (physician not required but can override), and level III (physician can contest). In contrast, Zekos (2021) rather distinguishes between three levels, that is, assisted, advisory, and autonomous AI.

The exact number of levels and their definition is not what matters here. What matters is that ethical concerns with respect to AI can be analyzed and classified according to such a conception of how AI can be used. Following Pineau's presentation at the 2019 congress of the Royal Society of Canada, AI can be seen to operate at four general levels, namely data gathering, prediction, prescription, as well as decision and action.

Level I - Data gathering AI, especially machine learning and deep learning, requires data. One of the most basic roles AI can perform is to accomplish this task and acquire data (e.g. Najjar-Ghabel, Farzinvash, and Razavi 2020). During this phase, algorithms can be used to fetch and store data (e.g. a smart watch monitoring heart rate, stress level, and respiration rate), but predictive analyses are not performed. Data gathering is considered as a part of assistive AI within the conception of the AMA, as well as at the basis of assisted AI in Zekos's (2021) terminology.

Level II - Prediction The second level consists in analyzing the acquired data through prediction analyses (e.g. a smart watch can predict one's race time). Prediction is implicit to assistive and augmentative AI as proposed by the AMA (i.e. detection is a prediction) and assisted AI as conceived by Zekos (2021). Predictive AI is meant to help individuals assess situations as well as possible choices and courses of action.

Level III - Prescription Prescriptive AI makes a recommendation based on the predictive analyses (e.g. a smart watch suggesting a training program). It corresponds to level I of autonomous AI as proposed by the AMA, and to advisory AI as proposed by Zekos (2021). Prescriptive AI is meant to suggest possible choices and actions in light of the predictive analyses made, while the choice remains a prerogative of the user.

Level IV - Decision and action The fourth level of AI corresponds to the implementation of autonomous choice and actions based on the predictions and prescriptions of the algorithm. It corresponds to Zekos's (2021) autonomous AI as well as the AMA's levels II-III of autonomous AI.

One interesting aspect of classifying AI through different levels of autonomy lies in the fact that ethical issues can be analyzed and classified accordingly, showing that there are issues intrinsic to each level that, in addition, combine themselves when going further up in the levels. Aside from the use of third party cookies, an interesting example illustrating ethical issues at the level of data gathering is the Duke Multi-Target Multi-Camera (MTMC) Dataset (Ristani et al. 2016; Satisky 2019; Tomasi 2019). In 2014, researchers at Duke University wanted to gather data meant to be used to train tracking algorithms as well as facial recognition software. Cameras where thus placed on campus to record individuals passing by. In the end, they were able to gather data involving 8 different cameras and approximately 2 000 individuals. However, some conditions imposed by the Institutional Review Board were not respected, and several issues can be highlighted in that example. First, the data collection took place outside, while it was supposed to be inside to ensure that only the images of individuals who gave their consent were captured. Second, there was not supposed to be any public access to the dataset, but the research team put it online, freely accessible to the general public and companies.

Two important values at play within that example are consent and privacy, the former superseding the latter. Indeed, whether or not there is a violation of privacy with respect to a specific state of affairs ultimately depends on the consent that is given regarding said state of affairs. Historically, and

more specifically with respect to research ethics, consent has been conceptualized through four fundamental characteristics (Faden and Beauchamp 1986).

Voluntary consent Consent needs to be voluntary in the sense that there needs to be no coercion at play. Coercion can be quite subtle and take various forms. At Duke University, for instance, a student might have needed to pass through a path where cameras were installed in order to get to class on time. In such a case, one could argue that such a student did not necessarily really consent to being filmed insofar as she had to pass through.

Informed consent Informed consent happens when individuals really know what they are consenting to. In Duke's case, for instance, even though students might have known they were being filmed (though it appears they were not), they could not anticipate that the dataset would be openly accessible online and that it would be used by companies to train algorithms that would then be used to track specific minorities in certain countries (as it was the case). Hence, even if one assumes that things would have been done properly in order to inform the students that they were being filmed, one could argue that an informed consent could not be obtained if all the relevant information regarding the potential usage of the data were not provided.

Competence to consent Requiring that a person is competent to give her consent is meant to ensure that her actions are consistent with her values and in her best interests. This competence can be understood on the grounds of the capacity to understand, to deliberate, and to make choices consistent with what one desires. The criteria used to evaluate this characteristic are generally taken to be legal rather than ethical (e.g., in the province of Quebec, Canada, this competence is governed by the Civil code).

Continuity Saying 'yes' to something before it happens does not imply that one will still agree once it has begun. As such, consent needs to be continuous throughout the data acquisition (and usage). Continuous consent will likely pose further challenges as AI capacities increase.

Consent is *the* ethical issue at the foundation of data gathering, and data gathering is an important part of AI. The reflections we currently have on regulations and guidance will determine how AI evolves. Article 18 of Bill C-27, for instance, states that businesses could collect individual's information without their consent *if it is a legitimate interest*:

> "**Legitimate interest** An organization may collect or use an individual's personal information without their knowledge or consent if the collection or use is made for the purpose of an activity in which the organization has a legitimate interest that outweighs any potential adverse effect on the individual resulting from that collection or use and (a) a reasonable person would expect the collection or use [...] and (b) the personal information is not collected or used for the purpose of influencing the individual's behavior or decisions."

Further, Article 18 leaves the ethical assessment of collecting data without consent to businesses by stating that it is their responsibility to "identify any potential adverse effect on the individual that is likely to result from the collection or use". Let us pause for a moment. What is a business' primary reason of existence? Making money, one might say. What tool do businesses usually use to reach that goal? Marketing. So marketing appears to be a legitimate interest of a business. But what is marketing if not an attempt at influencing people's choices? Food for thought.

The Duke example can be used to illustrate that ethical issues embed within one another as we go from one level of AI to another. As long as we are only having a dataset without using it (Level I), there are latent issues that do not yet have concrete empirical consequences insofar as they are, at that point, only violation of principles (e.g., privacy and consent). When we use these data, however, this latent violation can slowly turn into an actual prejudice. In Duke's case, tracking and facial recognition algorithms (Level II) where then used by some authorities to track and target minorities. Another example of a latent Level I ethical issue that becomes more important as we go from Level I to Level III is the presence of racist opinions, texts and comments (to say the least) on the web. Google Search autosuggestion, for instance, reinforces and normalizes biases by making predictive suggestions based on inappropriate data (e.g. associating gorillas to pictures of black individuals; Noble 2018). The important point to remember here is that there is a snowball effect of ethical issues when going from lower levels to higher levels. Ethical issues within Level I can develop and become harmful and highly prejudicial when algorithms are prescriptive or autonomous.

Back to the Greeks (Again)

It is worth emphasizing a parallel that can be made between ethics and mathematics. Mathematics, the queen of sciences (cf. the editorial in *Nature Computer Science* 2022), is generally taken as representative of a well-defined theory with well-defined methods and concepts. When discussing the historical context underlying the emergence of AI, Wooldridge (2021) wrote that Turing (and Church) had shown, by studying the halting problem, that "mathematics could not be reduced to following receipes (p.14)". But if mathematics, of all sciences, cannot be reduced to following receipes, how could ethics be? Some scholars see AI as applying to questions and problems that are taken to be decidable and have exact solutions, relying on the idea that there exist "precise and unambiguous methods for answering [these] questions (p.12)". Given such an understanding, ethical dilemmas cannot be solved by AI, for they have no exact solution. Ethical pluralism can be seen as the recognition of our own ignorance regarding what should be done. And this is the first step towards ethical choice: Knowing that there is no such thing as an ideal solution to a non-ideal situation.

Acknowledgments

This work was financially supported by the UQTR Research Chair in the Ethics of AI as well as by the *Fonds de Recherche du Québec* [2023-NP-310505]. It was further supported in part by funding from the Social Sciences and Humanities Research Council.

References

Adadi, A., and Berrada, M. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6:52138–52160.

Anderson, M., and Anderson, S. L. 2007. Machine ethics: Creating an ethical intelligent agent. *AI magazine* 28(4):15–15.

Anscombe, G. E. M. 1958. Modern moral philosophy. *Philosophy* 33(124):1–19.

Audard, C. 1999. *Anthologie historique et critique de l'utilitarisme*. Presses Universitaires de France.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The moral machine experiment. *Nature* 563(7729):59–64.

Awad, E.; Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2019. The thorny challenge of making moral machines: Ethical dilemmas with self-driving cars.

Bentham, J. 1834. *Déontologie ou science de la morale*. Les classiques des sciences sociales.

Biran, O., and Cotton, C. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8. 8–13.

Bjørgen, E. P.; Madsen, S.; Bjørknes, T. S.; Heimsæter, F. V.; Håvik, R.; Linderud, M.; Longberg, P.-N.; Dennis, L. A.; and Slavkovik, M. 2018. Cake, death, and trolleys: Dilemmas as benchmarks of ethical decision-making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 23–29.

CPT Editorial Panel. 2021. *Artificial intelligence taxonomy for medical services and procedures (Appendix S)*. American Medical Association.

Dancy, J. 2004. *Ethics without principles*. Oxford University Press.

De Cremer, D., and Kasparov, G. 2022. The ethical AI paradox: Why better technology needs more and not less human responsibility. *AI and Ethics* 2(1):1–4.

Dehaene, S.; Lau, H.; and Kouider, S. 2021. What is consciousness, and could machine have it? In von Braun, J.; Archer, M. S.; Reichberg, G. M.; and Sorondo, M. S., eds., *Robotics, AI, and Humanity: Science, Ethics, and Policy*. Springer. 43–56.

Etienne, H. 2022. When AI ethics goes astray: A case study of autonomous vehicles. *Social Science Computer Review* 40(1):236–246.

Faden, R. R., and Beauchamp, T. L. 1986. *A history and theory of informed consent*. Oxford University Press.

Faulhaber, A. K.; Dittmer, A.; Blind, F.; Wächter, M. A.; Timm, S.; Sütfeld, L. R.; Stephan, A.; Pipa, G.; and König, P. 2019. Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. *Science and Engineering Ethics* 25:399–418.

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5:5–15.

Gilligan, C. 1982. *In a different voice: Psychological theory and women's development*. Harvard University Press.

Harman, G. 1978. What is moral relativism? In Goldman, A. I., and Kim, J., eds., *Values and morals*, volume 13 of *Philosophical Studies Series in Philosophy*. Springer. 143–161.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Jørgensen, J. 1937. Imperatives and logic. *Erkenntnis* 7(1):288–296.

Kant, I. 1785. *Fondements de la métaphysique des mœurs*. Les classiques des sciences sociales.

Kaplan, A., and Haenlein, M. 2019. Siri, siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implica tions of artificial intelligence. *Business Horizons* 62(1):15–25.

LaCroix, T. 2022. Moral dilemmas for moral machines. *AI and Ethics* 2(4):737–746.

Li, J.; Zhao, X.; Cho, M.-J.; Ju, W.; and Malle, B. F. 2016. From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. Technical report, SAE Technical Paper.

Maclure, J. 2020. Context, intersubjectivism, and value: Humean constructivism revisited. *Dialogue* 59(3):377–401.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

Miller, B. 2021. Is technology value-neutral? *Science, Technology, & Human Values* 46:53–80.

Minister of Innovation, Science, and Industry. 2022. An act to enact the consumer privacy protection act, the personal information and data protection tribunal act and the artificial intelligence and data act and to make consequential and related amendments to other acts. House of Commons of Canada.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4):18–21.

Moore, G. E. 1959. *Principia Ethica [1903]*. Cambridge University Press.

Najjar-Ghabel, S.; Farzinvash, L.; and Razavi, S. N. 2020. Mobile sink-based data gathering in wireless sensor networks with obstacles using artificial intelligence algorithms. *Ad Hoc Networks* 106:102243.

Nascimento, A. M.; Vismari, L. F.; Queiroz, A. C. M.; Cugnasca, P. S.; Camargo, J.; and de Almeida, J. 2019. The moral machine: Is it moral? In *Computer Safety, Reliability, and Security: SAFECOMP 2019 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Turku, Finland, September 10, 2019, Proceedings 38*, 405–410. Springer.

Noble, S. U. 2018. *Algorithms of oppression*. New York University Press.

Noothigattu, R.; Gaikwad, S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. 2018. A voting-

based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ogien, R. 2007. *L'éthique aujourd'hui: Maximalistes et minimalistes*. Gallimard.

Parfit, D. 1984. *Reasons and Persons*. Oxford University Press.

Pellé, S., and Reber, B. 2016. *Éthique de la recherche et innovation responsable*, volume 2. ISTE Group.

Peterson, C., and Broersen, J. 2024. Understanding the limits of explainable ethical AI. *International Journal on Artificial Intelligence Tools*.

Peterson, C., and Hamrouni, N. 2022. Preliminary thoughts on defining $f(x)$ for ethical machines. *The International FLAIRS Conference Proceedings* 35.

Pineau, J. 2019. The rise of AI: Who is making decisions about our health? In *Science, Trust and Democracy in the Digital Age, September 19-20, Ottawa*. Royal Society of Canada.

2022. Mathematics, the queen of sciences. *Nature Computer Science* 2(409).

Rawls, J. 1987. *Théorie de la justice*. Éditions du Seuil.

Rawls, J. 2005. *Political liberalism*. Columbia University Press.

Raz, J. 1999. *Practical Reason and Norms*. Oxford University Press.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, volume 9914 of *Lecture Notes in Computer Science*, 17–35. Springer.

Ryan, M. 2020. In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics* 26:2749–2767.

Satisky, J. 2019. A Duke study recorded thousands of students' faces. Now they are being used all over the world. *The Duke Chronicle*.

Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3):417–424.

Sen, A., and Williams, B. 1982. *Utilitarianism and Beyond*. Cambridge University Press.

Sparrow, R. 2004. The Turing triage test. *Ethics and Information Technology* 6:203–213.

Tomasi, C. 2019. Letter: Video analysis research at duke. *The Duke Chronicle*.

Torrance, S. 2011. Machine ethics and the idea of a more-than-human world. In Anderson, M., and Anderson, S. L., eds., *Machine ethics*. Cambridge University Press. 115–137.

Turri, J. 2010. On the relationship between propositional and doxastic justification. *Philosophy and Phenomenological Research* 80(2):312–326.

UNESCO. 2021. *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific and Cultural Organization.

Van de Poel, I., and Royakkers, L. 2011. *Ethics, technology, and engineering: An introduction*. John Wiley & Sons.

van den Hoven, J. 2010. The use of normative theories in computer ethics. In Floridi, L., ed., *The Cambridge Handbook of Information and Computer Ethics*. Cambridge University Press. 59–76.

Wang, H.; Khajepour, A.; Cao, D.; and Liu, T. 2020. Ethical decision making in autonomous vehicles: Challenges and research progress. *IEEE Intelligent Transportation Systems Magazine* 14(1):6–17.

Weinstock, D. 2017. Compromise, pluralism, and deliberation. *Critical Review of International Social and Political Philosophy* 20(5):636–655.

Wolf, S. 2021. Differences between natural and artificial cognitive systems. In von Braun, J.; Archer, M. S.; Reichberg, G. M.; and Sorondo, M. S., eds., *Robotics, AI, and Humanity: Science, Ethics, and Policy*. Springer. 17–27.

Wooldridge, M. 2021. *A brief history of artificial intelligence: What it is, where we are, and where we are going*. Flatiron Books.

Yampolskiy, R. V. 2013. What to do with the singularity paradox? In Müller, V. C., ed., *Philosophy and theory of artificial intelligence*. Springer. 397–413.

Zekos, G. I. 2021. Artificial intelligence governance. In *Economics and Law of Artificial Intelligence: Finance, Economic Impacts, Risk Management and Governance*. Springer. 117–146.