Preliminary Thoughts on Defining $f(x)$ for Ethical Machines

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

Abstract

There is a growing literature in machine ethics attempting at creating ethical machines through AI and machine learning. Although many concerns with respect to such attempts have been raised, including the difficulties regarding the gathering of relevant contextual information as well as solving ethical dilemmas, it appears that many fundamental ethical notions have been overlooked in the implementation of normative theories to machines. This paper provides a preliminary analysis of important aspects that need to be taken into account in the attempt of defining so called ethical machines.

On computable ethical behavior

With the increasing place artificial intelligence (AI) is taking in our lives, as well as a rapidly growing reflection on what AI is, what it should be, and what it should do or be used to, scholars are assuming that ethical behavior for machines is possible and that we should thrive to reach such standards (cf. Etzioni and Etzioni 2017; Wallach and Allen 2009). Despite a tremendous amount of literature on formal models of ethical behavior (see Peterson 2016), ranging from models of reasoning to action logic and passing by multi-agent systems, and despite many red flags raised with respect to the practical and theoretical limits of automating ethical reasoning (Dignum 2019; Peterson 2020), scholars are still pursuing the idea that moral machines are not only possible, but in a sense inevitable (e.g., Anderson and Anderson 2011). Yet, the idea of an *ethical machine* is not trivial, and it is important to clarify in which sense (if at all) a machine could be ethical. In addition to the considerations regarding the ethical implications such a machine would have (e.g., theorizing of moral agency), one further needs to understand how ethics would be encoded within the machine, and how that machine would actually (physically) work.

The objective of this paper is to highlight problems related to the idea of ethical machines by grouding ethical machines into the ethics literature and examining the constraints surrounding the programmation of such machines. Doing so will allow us to point out specific elements that should be taken into account in the attempt at developing such ethical machines. In what follows, we begin by analyzing the relationship between machine ethics and the singularity hypothesis, which will bring us to a discussion of ethics and the limitations of normative theories as well as how these normative theories are applied to machine ethics. Agency and responsibility will be discussed and presented as red flags to autonomous moral agents, and then we will analyze what it would really mean to define a function meant to guarantee ethical behavior for machines. We conclude by discussing the ramifications of such an ethical function and highlight important aspects to consider in the attempt of defining it.

On the possibility of an ethical singularity

Reflections on ethical AI can be divided within two broad categories: Those assuming that machines can by themselves be ethical and see machines as (so called) autonomous moral agents (AMA), and those seeing machines and AI (i.e., the use of automation and optimization techniques based on algorithms and advanced statistical techniques to pursue specific goals) as being possible targets of moral evaluation (cf. Johnson in Anderson and Anderson 2011). Among the former are those reflecting upon ethical machines from the perspective of the singularity, that is, the possibility that some form of consciousness or self-awareness would supervene on the physical properties and complexity of the machine. Without going as far as to discuss specific issues related to singularity and its affiliated problems (cf. Anderson and Anderson 2011; von Braun et al. 2021; Dubber, Pasquale, and Das 2020; Floridi 2010; Müller 2013), it is important to emphasize that the idea of an ethical singularity is fundamentally inconsistent with ethics and that the mere idea of an autonomous moral agent is deeply flawed, at least from an ethics (and, dare we say, rational) perspective. Indeed, there is a profound inconsistency and considerable theoretical limitations surrounding the idea that an autonomous machine could be *ethical*. While singularity and the idea that conscience could emerge from a machine rely on a deterministic conception of the mind, with Turing's (1950) *skin-of-an-onion* analogy as emblematic of the mechanical understanding of human thoughts, ethics relies on an assumption that has been opposed to determinism throughout the history of rationality: Freedom to act. That moral responsibility requires freedom to act follows from the fact that it would be irrelevant to blame someone for events that were not in one's power to change. To be responsible for an event is to have made choices that eventually lead to that event. In a nutshell, the problem lies in the fact that moral responsibility requires freedom to act, whereas determinism implies the negation of free will and, incidentally, of moral responsibility. The fundamental inconsistency pertaining to the idea of an ethical singularity therefore lies in the reconciliation of the deterministic presupposition that every empirical phenomenon is reducible to a series of causes and effects with the idea that individuals can themselves choose right from wrong (i.e., that ethical behavior is possible).

This opposition (cf. Campbell, O'Rourke, and Shier 2004; Cohen and Trakakis 2008; Sinnott-Armstrong 2014), also known as the debate between compatibilism (i.e., free will and moral responsibility can be reconciled with the causal and deterministic understanding of the world) and incompatibilism (i.e., determinism implies that we do not have the free will necessary to account for moral responsibility), has diverged in many fields including philosophy of mind, social psychology and quantum mechanics and has incidentally evolved into a metaphysical debate regarding the physical structure of the world and the essence of acting. As empiricists and pragmatists, we do not see any incentive to dwell into metaphysics and assume the possibility of an ethical singularity: The world can be understood without any commitment to the belief that determinism is true, whereas our actions and choices cannot be objectively evaluated without ethics. From a pragmatic standpoint, ethics should be favored over determinism (i.e., in the absence of any concrete evidence favoring determinism, there is a pragmatic choice to focus on the possibility of objective and rational ethical behavior). But this position on determinism versus responsibility is not the important point here. What is important is that advocates of the singularity hypothesis should really be thinking about the ramifications of their assumptions from an ethical perspective. For if the singularity implies that objective and rational ethical behavior is impossible, then one should be quite careful with attempts at creating the singularity and focus on concrete theoretical and technical security measures (Brundage 2014). Notwithstanding that we do not think that real (substantial) autonomous moral agents are possible, it should be emphasized that our analysis in what follows is actually independent of whether the singularity is possible or not. Whether or not one is able to reconcile the singularity hypothesis with the fundamental assumptions underlying ethics, there are key elements to consider in the attempt to define computable functions that could be used to determine ethical behavior.

What is ethics?

To understand how a machine could be ethical, one needs to understand what ethics is all about. Ethics is more than mere consequentialism. It also goes beyond following the rules. Ethics is a systematic reflection on the values, principles and norms that should guide our actions and behavior. On the assumption that moral relativism (i.e. a theory postulating that ethical evaluations depend upon one's perception or conception of ethical norms and values, which cannot objectively be evaluated) can be refuted and that objective ethical judgment is possible, philosophers have developed normative theories in order to expose how and why we can provide objective moral evaluations. The most studied (families of) theories are utilitarianism/consequentialism, deontology, virtue ethics and, since the last thirty years, care ethics.

Utilitarians such as Bentham and Mill argued that morality is grounded on the fact that we are all sentient beings: Everybody has an interest in seeking pleasure and avoiding pain. As such, Bentham advocated a direct (or total) form of utilitarianism, where ethical actions are those that maximize utility (here understood as the overall sum of pleasures in balance with the overall sum of pains), whereas Mill argued in favor of an indirect form of utilitarianism, where an action is required when it follows a rule that maximizes overall well-being (which allows to avoid the objection that utilitarianism lead to the sacrifice of minorities). Utilitarianism eventually led to a broader class of theories dubbed consequentialism (Sen and Williams 1982). Consequentialist theories advocate that one's actions should be evaluated with regard to the values they promote (e.g., justice, autonomy, fairness), either by maximizing the outcome (i.e. maximizing consequentialism; e.g. Bentham) or by requiring the satisfaction of a fixed threshold (i.e. satisficing consequentialism; e.g. Slote and Pettit 1984). Actions are thus evaluated on the grounds of the values they reach through their consequences, bearing in mind the scope of its effects (e.g., number of persons affected), the length of the effects, whether the consequences are likely or unlikely, and whether the action has side effects (good or bad). Important characteristics of consequentialist theories are impartiality (i.e., everyone should be considered equality within the evaluation), universality (i.e. everyone's interests should be considered), and aggregation (i.e. quantification of costs/benefits).

Deontological ethics includes theories advocating that actions should be accomplished when they are required by a norm, notwithstanding their consequences, the context, the agent's characteristics, or her moral psychology and emotions. Kant, for instance, defended that morality emerges from one's rationality and capacity to act in accordance to universal principles. As rational beings, individuals have the autonomy to recognize that human dignity is an end in itself, and have the capacity to choose to act towards that end. Acknowledging human dignity implies the recognition of reciprocity (which is also a criterion for distinguishing between moral agents [capable of ethical behavior] and patients [that should be considered within the ethical evaluation]), meaning that one should always consider others as ends in themselves, and never as mere means to an end. For Kant, human life has a sacred value and everyone should be considered as equals, and it does not suffice to respect the rules to act ethically: One has to be willing to act in accordance to these rules (Kant's *good will*), which emphasizes the importance of one's source of motivation and intentions in the ethical evaluation of one's action. Intention is an intrinsic component to ethical action, and one needs to want to act morally in order to actually act morally, independently from one's natural inclinations, instincts, or emotions.

Virtue ethics began with Aristotle's writings (later revisited by Anscombe 1958). Advocates of virtue ethics insist on agent's characteristics, virtues and vices and try to determine the circumstances under which an individual is good rather than the particularities of a good action. Insisting on an individual's characteristics rather than on actions or rules, it advocates that good actions are those that would be performed by virtuous agents and, as such, that one needs to concentrate on what makes a person good. Based on character traits emerging from processes of socialization, education, environmental contexts and genetics, individuals are predisposed to act in specific ways. Accordingly, virtue ethics insists on developing one's predispositions to act virtuously, for instance by learning self-control, moderation, compassion or generosity (types of virtues vary depending on authors; e.g. contemporary aristotelians value courage, magnanimity and political participation).

On the limits of normative theories

The main ethical traditions suffer from well-known limitations. For instance, when the cost vs benefice analysis is fitting, consequentialism allows to sacrifice human life and tend to favor the interest of the majority, whereas deontological ethics does not consider the consequences resulting from the application of absolute principles and struggles with conflicting obligations. One notable objection to the main ethical currents and that is especially relevant to machine ethics stems from the work of moral psychologist Carol Gilligan (1982), who realized that women, when confronted to hypothetical ethical dilemmas, did not reason on the grounds of the two main principilist ethical traditions. Instead of analyzing dilemmas in terms of rights and duties or respecting the utility principle, women were rather concerned about the well-being and emotions of the individuals and their judgment varied depending on contextual characteristics. As opposed to consequentialism and deontology, which advocate general principles that can be used to distinguish right from wrong, care ethics emphasizes the relevance of contextual specificities in analyzing moral dilemmas and highlights the importance of caring and showing sincere concern for individuals' well-being (both for oneself and others). Care ethics has thus been developed in reaction to what can be seen as an androcentric bias in the historically dominant ethical traditions, which rely on rationality, impartiality, and universalism rather than emotions, needs, and concern for others.

While normative ethical theories were initially meant to provide an answer to moral relativism and provide objective and rational foundations for ethical judgment, the fact that there is no universally endorsed moral theory is often seen as reopening the door to the relativist objection. Indeed, some see the fact that all these normative theories contradict themselves as an argument in favor of the idea that there is no such thing as a true ethical theory and, incidentally, as an argument for ethical relativism. While we agree that there is no such thing as *the* true normative theory, it should be emphasized, however, that this does not necessarily lead to ethical relativism, and that this diversity of ethical theories and principles should really be seen as a strength rather than a weakness. Ethical pluralism amounts to the idea that there is no aspect that is intrinsically superior to others in the ethical evaluation of a situation. Hence, it is always relevant to evaluate the person, her actions, her intentions, the rules, the values at stake (including needs and well-being), the consequences of her action as well as the specifics of the situation to provide an ethical assessment. Ethical pluralism can provide an appropriate answer to moral relativism, as long as one accepts that ethical problems do not have a unique solution. Instead of arguing that one ethical theory is true or well justified whereas the others are not, which would imply many conceptual difficulties, ethical pluralism recognizes the plurality of reasonable ethical positions and insists on reaching compromises to solve ethical dilemmas. For instance, ethical dilemmas can be resolved through discussion (Weinstock 2017) in different (though often incompatible) ways, and the morality of our actions and choices can be evaluated on the grounds of the reasons invoked. Resolving ethical dilemmas in concrete situations necessarily requires the sacrifice of ethical principles or values (Weinstock 2017), and this sacrifice needs to be evaluated in light of the specificity of the situation (Peterson 2020). Ethical pluralism not only acknowledges a multiplicity of ethical theories and values but also recognizes the fact that many dimensions can simultaneously be subjected to ethical scrutiny. For instance, one can evaluate the intention and character traits of an individual, the values, the norms in place meant to guide one's actions, or the consequences and risks of one's actions, which are all different aspects advocated by the main ethical currents. By recognizing that there is a space of possibly reasonable and acceptable ethical solutions rather than a unique ethical truth, one can navigate through ethical theories and weight principles and values given contexts and situations.

Ethical behavior for machines

Despite the limits of ethical theories and based on the idea that ethical rules can properly be encoded within machines, scholars are pursuing the idea of ethical machines by trying to implement normative theories in machine development (e.g., Anderson and Anderson 2007; Muehlhauser and Helm 2012). Implementation of ethical behavior is conceived either from a top-down (i.e. directly coding ethical principles within the machine), a bottom-up (i.e. using machine learning to be able to imitate actual human reasoning and behavior), or a hybrid approach (see Tolmeijer et al. 2020 for a comprehensive overview). There are known limits, however, to such attempts. Tolmeijer et al. (2020) identified practical problems such as rule selection and conflict resolution, the determination of possible consequences (including the size, the scope, and the probability of the effects; the definition of utility measures, and computational costs), and applying notions of personality and character traits to machines to assess ethical behavior. Brundage (2014) further identified limitations for ethical machines, including problems pertaining to i) insufficient knowledge or resources, which can result in type I (i.e. generating an exception to a rule when it should not be the case) and type II (i.e. not generating an exception to a rule when it should be the case) errors, ii) resolving moral dilemmas, as well as iii) the inappropriate training of algorithms and the biases in the data.

Agency, responsibility and liability

Assuming that machines could act ethically, and that one made a mistake, would the machine be responsible? Thinking of AI as a possible bearer of rights and duties has its share of problems (cf. Kingwell in Dubber, Pasquale, and Das 2020), including the difficulty of appropriately defining what a *person* is. Although it might be of conceptual interest to reflect upon the prerequisites as well as the ramifications of conceiving machines and AI as persons, it is important to emphasize that machines consist of physical parts and lines of code and that thinking of machines as potential right-holders is not a trivial matter (cf. Basl in Dubber, Pasquale, and Das 2020). While some authors see the possibility of the singularity hypothesis as a reason to pursue academic research on the possible integration of AI as a social agent (e.g. Gips in Anderson and Anderson 2011), others, like ourselves, emphasize the importance of focusing on concrete issues that have dire consequences on human lives, including a proper framework for responsible AI, controlling the social impact of AI (e.g., discrimination, inequities) as well as the integration of ethical considerations within technological developments.

Johnson (in Anderson and Anderson 2011) argued that computers are not independent autonomous moral agents insofar as they do not have internal desires, beliefs, or intentional states and, accordingly, they should not be considered as moral agents in themselves, independent from human behavior (including choices and programmation). Following Sullins (in Anderson and Anderson 2011), the interpretation of moral agency relies on notions such as autonomy, intentionality, and responsibility. As Johnson pointed out, it is important to distinguish between moral agency (i.e. satisfying the conditions required for possible ethical behavior) and moral patients (i.e. something that can be the target of an ethical assessment; e.g., guns should be restricted; offshore accounts are unjust; facial recognition software favors racial discrimination). AI and machines can be the subject of moral evaluations, but they should not be considered as moral agents insofar as they do not properly satisfy prerequisites for moral agency.

It is noteworthy that from a conceptual and ethical perspective, all the aforementioned prerequisites for agency are intertwined. For instance, autonomous action requires i) intention and volition (i.e. internal mental states), ii) comprehension (including the motivation, the end, the means, and possible consequences), and iii) independence from external control (Faden and Beauchamp 1986). Similarly, responsibility presupposes the capacity to choose and act as well as an intentional internal state. To put it differently, autonomy requires intent (i.e. intentionality), and responsiblity requires autonomy. (i.e. intent and a capacity to freely choose and act). In line with this, Dignum (2019) emphasizes that ethical action is characterized by the possibility of choice as well as the agent's internal motivation to act ethically.

One should be quite careful when trying to interpret a machine as an autonomous moral agent. This is not to say that AI and machines cannot be the object of an ethical evaluation. The point is that moral agency requires moral responsibility, and responsibility includes several dimensions such as blameworthiness, liability, and accountability (Pellé and Reber 2016). Moral agents act in non-ideal situations (i.e., in which conflicting principles and values are at play) bearing in mind that they will be accountable for their choices. As long as it does not make any sense to consider a machine as accountable (e.g., think of reprimanding or punishing a machine), one should refrain from seeing such a machine as responsible and, incidentally, as a moral agent.

Would a machine care?

Beside the problems and challenges pertaining to the implementation of normative theories, it should be emphasized that machine ethics is primarily conceived and implemented in relation to consequentialism, deontological ethics, and hybrid approaches between these two (cf. Tolmeijer et al. 2020). Looking thoroughly at Tolmeijer et al.'s (2020) survey, one will see no mention of care ethics or ethical pluralism, with a "surprisingly low percentage of authors" considering the application of multiple ethical theories. One important aspect of ethical reasoning and behavior that does not seem to be discussed within the machine ethics literature (aside from the superficial and instrumental analysis, detection and modeling of emotions) is the role played by emotions, sympathy and empathy within ethical assessments. When evaluating consequences and repercussions of actions, one needs to be able to put aside one's feelings and desires and try to put oneself in someone else's position in order to really understand the implication of one's action in others' lives. Ethical behavior is only possible when one is able to objectively evaluate one's own beliefs, desires and preferences and put them in balance with the ethical principles and values at stake in light of the specific characteristics of the situation. Ethical pluralism dictates that ethical behavior requires the consideration of emotions, needs, and concern for others within the ethical assessment of a situation. As long as a machine can not sincerely and deeply care, one should not consider machines as moral agents.

Defining $f(x)$

Although scholars are working on programming machine ethics (cf. Tolmeijer et al. 2020; Pereira and Saptawijaya 2016), there are basic programming notions with deep ethical ramifications that seemed to have been overlooked by the community, namely the basic characteristics of a computer function. As Johnson (in Anderson and Anderson 2011) wrote, what machines "do is receive input and transform the input into output". A pragmatic analysis of ethical machines thus requires that we reflect upon the type of input, the transformation process, and the type of output that an ethical function would require. Dignum (2019) sees an ethical function as taking as input an action a, characterized by a set of preconditions (specifying when the action can be accomplished) as well as some form of ethical value, and a context c, specifying the conditions under which the agent needs to act. Though she points out that there are practical limitations rendering impossible the appropriate implementation of such a function (e.g., that all actions available given a specific context can be determined, valued, and ranked ethically; this can also partly be conceived as the frame problem, cf. McCarthy and Hayes 1969), such an understanding of what $f(x)$ would look like is an oversimplification of what ethical assessment is. (The following notation is to illustrate and should not be understood as a formal definition.)

The main point of ethical pluralism is not only that there is a diversity of ethical principles, theories and values that need to be considered in the analysis of ethical dilemmas, but also that there is a plurality of things (e.g., actions, norms, persons, consequences, artifacts, values, etc.) that can be the object of an ethical evaluation. Thus, right from the start, a computer function meant to regulate ethical behavior would need to consider a variety of input types. Of course, we expect a good programmer to divide the task into subfunctions meant to deal with specific types of inputs, but this would be an oversimplification insofar as all these subfunctions are intertwined within the ethical assessment of a situation, and how they relate to each other (e.g., morality of an act given the rules and its consequences), even if this relationship could be predicted, would only yield a variety of possibly ethical outcomes depending on the norms, values and principles that are prioritized and on the context variables that are deemed important (cf. frame problem). How would these subfunctions work? For instance, we might expect that a function taking as inputs norms (n), contexts (c), and priority rules (p) would output the correct norm to apply to a situation. Call this function $N(n, c, p)$. This might seem simple enough, but it does not take into account that deciding which norm applies is in itself subject to an ethical evaluation (i.e. n and p need to be evaluated ethically). Depending on the evaluation of n and p, there is a variety of $N_1, \ldots, N_i$ to take into account. For instance, we would have different N_i from a deontological or a consequentialist standpoint. So which one should we choose? A clever programmer would perhaps say that the solution is also quite simple: One simply has to define another function $F(N_1, \ldots, N_i)$ that provides an ethical assessment (e.g., an ordering) of these normative functions and that output the N_i that should be retained. Recursive thoughts, right? But wait: How can we determine which N_i should be retained? There is a variety of $F_1, \ldots, F_i$ to choose from (depending on the principles to which one appeals to define the ordering), and this also requires an ethical assessment. Machine ethics therefore falls under the scope of Moore's (1959) open question argument: One can always ask whether the output of an ethical function (or the function itself) is the correct one. For every function that is defined, one can always wonder whether this function is really ethical or whether there are better alternatives. Put differently, any ethical function f can be the input of a more general function meant to determine whether f is indeed ethical: There is no absolute ethical function that cannot be itself subject to an ethical evaluation. Accordingly, the open question argument implies that ethics cannot be functionally defined. Apart from these considerations, there is another basic problem with the idea of a general ethical function: What would be the type of the output of such a general ethical function? Would the function be binary, with outputs good/bad, just/unjust, ethical/unethical? Would it provide an ordering of possible alternatives? Would the action with ethical value closest to 1 be retained? Scholars seem to naively conceive such an ethical function as something that would provide us with *the* action to accomplish or the *correct* action to do in specific situations. But this is misleading. At best, such a function would provide us with a set of ethical possibilities. Trying to impose an ordering on this set would require one to leave the realm of ethical pluralism and endorse a specific normative theory, and normative theories are known to fail.

Would a machine be justified to act?

What does it mean to be *ethical*? There is a distinction to be made between what ethics is, and how ethical concepts are used (also known as folk morality). People use *ethical* as a substitute for the *correct* or the *right* thing to do. They see an ethical action as an action that is *justified* given ethical values and principles, that is, as an action that had to be done given the circumstances. Scholars in machine ethics seem to have a tendency to assume that unequivocal outputs can be obtained from an ethical assessment. This, however, would be misunderstanding what ethics is all about. Consider Power's (in Anderson and Anderson 2011) example of a resolution of nonmonotonic reasoning with default rules:

> "An ethics example might be the default rule 'Don't kill the innocent.' The defeating conditions might be 'unless they are attacking under the control of some drug' or 'except in a just war,' and so on."

The basic idea behind that example is that the general rule 'It is forbidden to kill the innocent' can be defeated under specific circumstances, for instance in cases of self-defense. Consider the rule as a declarative sentence. In the context of self-defense, this means that 'It is forbidden to kill the innocent' is false, from which 'It is permitted to kill the innocent' can be derived through vary basic inferential principles (Peterson 2016). Put differently, one might say that an individual is *justified* to kill under the circumstances of self-defense. People tend to see ethical justification as entitlement. One is *justified* to act insofar as one has the *right* to act in such a way. One *must* act in such a way because it is *the* thing to do. However, this understanding relies on a misapprehension of what ethics is as well as how human behavior needs to be understood in sub-ideal situations. Ethical behavior emerges with ethical dilemmas (i.e., conflicting ethical values and principles) in a non-ideal world (cf. Jones and Pörn 1985). In this context, one must do as best as one could given the circumstances, bearing in mind that in the end there will be an upside (benefice) and a downside (costs). From a conceptual standpoint, ethical justification refers to the reasons one appeals to to support the choice or the action that is made. But this justification should not be understood as entitlement. These reasons do not give one the *right* to act in such a way. They do not imply that the action conforms to an ideal of justice. An action that can be ethically justified is not necessarily *the* action that *has to be* accomplished. Rather, ethical justification needs to be understood as an excuse: Given the circumstances, one can be excused to have behaved in such a way, or made such a choice. The action or choice will be ethical not because it was the one to do, but because one can understand (even if one disagrees) why someone acted in such a way given the values at stake. We can understand why one would kill under the circumstances of self-defense and we can excuse that behavior, but this does not mean that one is entitled (or even allowed) to do so. Ethical pluralism insists on the fact that there is no such thing as *the* (one and only) right thing to do. There is a variety of choices with respect to ethical dilemmas that, as rational agents, we can understand, and the agent making the choice needs to understand why the choice was made (i.e., the ranking of the principles and values that support the choice), the consequences of the choice and, more importantly, the agent needs to keep in mind that she will be accountable and liable for her actions. We, as humans, have to live with our mistakes and mishaps. We are responsible for our actions, and this makes us reevaluate our ordering of values and principles as well as the weighting of potential consequences in the ethical evaluation of a situation.

Closing thoughts

We can only but agree with Dignum and Johnson that although there is ethics and responsibility to be integrated to technological developments, real ethics and real responsibility relies upon managers, developers, users, and governments. *We* are responsible. Machines and AI should not be considered as moral agents capable of ethical behavior. It does not suffice to be able to imitate human reasoning to think ethically. For a machine to be intrinsically ethical, it would require that it is responsible and that it cares. It would require a capacity for empathy as well as a constant monitoring and reevaluation of its own values, rules, and biases, in balance with the context and its relevant ethical aspects. As humans, we struggle to do so. But this is okay, as ethics does not mean actual perfection. On the contrary, one can only aspire to reach ethical perfection, and, as Aristotle argued, it can be quite difficult to establish the precise threshold allowing to discriminate between acceptable and blamable behavior. We, as humans, fail to act as we would in an ideal ethical world, where no ethical dilemmas would arise. But we thrive to, and it is the attempt that counts.

References

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

Anderson, M., and Anderson, S. L. 2011. *Machine ethics*. Cambridge University Press.

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

Brundage, M. 2014. Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence* 26(3):355–372.

Campbell, J. K.; O'Rourke, M.; and Shier, D., eds. 2004. *Freedom and Determinism.* A Bradford Book.

Cohen, D., and Trakakis, N. 2008. *Essays on Free Will and Moral Responsibility.* Cambridge Scholars Publishing.

Dignum, V. 2019. *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer.

Dubber, M. D.; Pasquale, F.; and Das, S., eds. 2020. *The Oxford Handbook of Ethics of AI.* Oxford University Press.

Etzioni, A., and Etzioni, O. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 21(4):403–418.

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

Floridi, L., ed. 2010. *The Cambridge handbook of information and computer ethics*. Cambridge University Press.

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

Jones, A. J. I., and Pörn, I. 1985. Ideality, sub-ideality and deontic logic. *Synthese* 65(2):275–290.

McCarthy, J., and Hayes, P. J. 1969. Somephilosophical problems from the standpoint of artificial intelligence. In Metzer, B., and Michie, D., eds., *Machine Intelligence 4*. Edinburgh University Press. 463–502.

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

Muehlhauser, L., and Helm, L. 2012. The singularity and machine ethics. In Eden, A.; Moor, J.; Søraker, J.; and Steinhart, E., eds., *Singularity Hypotheses*, The Frontiers Collection. Springer. 101–126.

Müller, V. C., ed. 2013. *Philosophy and theory of artificial intelligence*. Springer.

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

Pereira, L. M., and Saptawijaya, A. 2016. *Programming machine ethics*, volume 26. Springer.

Peterson, C. 2016. *De la logique des obligations, des permissions, et des interdictions: De von Wright à aujourd'hui*. Presses de l'Université de Montréal.

Peterson, C. 2020. How to act? Reasoning with conflicting obligations. In *Proceedings of the Thirty-Third International FLAIRS Conference*. Association for the Advancement of Artificial Intelligence.

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

Sinnott-Armstrong, W., ed. 2014. volume 4: Free will and moral responsibility. Bradford Books.

Slote, M., and Pettit, P. 1984. Satisficing consequentialism. *Proceedings of the Aristotelian Society* 58:139–176.

Tolmeijer, S.; Kneer, M.; Sarasua, C.; Christen, M.; and Bernstein, A. 2020. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)* 53(6):1–38.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460.

von Braun, J.; Archer, M. S.; Reichberg, G. M.; and Sorondo, M. S., eds. 2021. *Robotics, AI, and Humanity: Science, Ethics, and Policy*. Springer.

Wallach, W., and Allen, C. 2009. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

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