Further Thoughts on Defining $f(x)$ for Ethical Machines: Ethics, Rational Choice, and Risk Analysis

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

Abstract

There is a tendency to anthropomorphize artificial intelligence (AI) and reify it as a person. From the perspective of machine ethics and ethical AI, this has resulted in the belief that truly autonomous ethical agents (i.e., machines and algorithms) can be defined, and that machines could, by themselves, behave ethically and perform actions that are justified from a normative standpoint. Under this assumption, and given that utilities and risks are generally seen as quantifiable, many scholars have seen consequentialism (utilitarianism) and rational choice theory as likely candidates to be implemented in automated ethical decision procedures, for instance to assess and manage risks as well as maximize expected utility. Building on a recent example from the machine ethics literature, we use computer simulations to argue that technical issues with ethical ramifications leave room for reasonable disagreement even when algorithms are based on ethical and rational foundations such as consequentialism and rational choice theory. By doing so, our aim is to illustrate the limitations of automated behavior and ethical AI and, incidentally, to raise awareness on the limits of so-called ethical agents.

On autonomous ethical choices

As the scientific literature on machine ethics and ethical artificial intelligence (AI) keeps growing, and although many red flags have been raised casting doubts on the mere possibility of defining truly autonomous ethical machines (Peterson and Hamrouni 2022), scholars are attempting to define algorithms that would allow machines to perform autonomous ethical choices. Tolmeijer et al. (2020) recently provided a thorough survey of not only how machine ethics is implemented (e.g., ethical theories such as deontology, consequentialism, or virtue ethics; top-down, bottom-up and hybrid approaches; technology type including hardware as well as logical, statistical, and probabilistic reasoning), but also of the shortcomings of these implementations, pointing out well-known theoretical (e.g. variety of incompatible ethical theories), practical (e.g. conflicts between rules) and technical (e.g. computation time) limitations of the current approaches as well as the challenges future approaches will face (see also Brundage 2014; Dignum 2019; Johnson 2011). Yet, despite these concerns and limitations, scholars are pursuing the idea that truly autonomous ethical machines can be defined (e.g., Anderson and Anderson 2007; Muehlhauser and Helm 2012). To exemplify, Anderson and Anderson (2007), building on Moor's (2006) distinction between implicit (i.e., ethical constraints and principles imposed beforehand in the programming and design), explicit (i.e., ability to represent ethics and make choices on the grounds of this knowledge) and full ethical agents (i.e., ability to make explicit ethical judgment and justify them), argue that the ultimate goal of machine ethics is to create (at least) explicit ethical agents that would be able to make autonomous choices based the representation of some ethical theory. Furthermore, they see full ethical agents as within the reach of machine ethics, arguing that one of their benefit would be their ability to "make correct ethical judgments" and "explain why a particular [choice] is either right or wrong by appealing to an ethical principle" (i.e., they would be able to provide reasonable justifications for their choices; see Moor 2006). Hence, their work, which focuses on decision making and ethical choice, is motivated by the idea that machine ethics can create not only explicit ethical agents, but that it can further create full ethical agents.

In reaction to this trend in machine ethics, Peterson and Hamrouni (2022) argued that such attempts are misconstrued, for there is no such thing as an ethical choice without (among other things) responsibility, compassion and compromise (see also De Cremer and Kasparov 2022; Dignum 2019, 2021). Building on Moore's (1959) open question argument, they further argued that ethics cannot be functionally defined (i.e., one cannot define 'the' ethical function $f(x)$ that would always yield 'the' correct answer), for one will always be legitimately able to question whether $f(x)$'s output is indeed 'the' correct choice, or whether $f(x)$ is indeed correctly defined. Put differently, the question of whether a choice made by a machine or an algorithm was indeed 'the' ethical one will always be open and subject to reasonable disagreement. To be explicit, Peterson's and Hamrouni's (2022) analysis relies on ethical pluralism, which recognizes a plurality of reasonable (and incompatible) ethical positions as well as the fact that there is nothing in an ethical evaluation that should always dominate other aspects (see also Maclure 2020; Weinstock 2017). From the perspective of ethical pluralism, there is no such thing as

'the' correct ethical choice: There is only a space of (mutually exclusive) excusable or understandable choices that are available given specific circumstances (on this point, see also Dancy 2004). And this is precisely why ethical choice implies (among other things) responsibility, because a (real) ethical agent is one that will be held responsible and accountable for the choice that has been made in a specific situation in order to reach a compromise, and where other reasonable (or ethical) choices were available.

The aim of the present paper is to pursue Peterson's and Hamrouni's endeavor and exemplify why the very idea of autonomous ethical (moral) agents is misconstrued by showing that ethical pluralism and reasonable disagreement emerge even from a technical standpoint during the construction of decision procedures, casting doubts on whether these autonomous agents were ethical in the first place. Reflecting on the implementation of ethical choice from an applied perspective, our aim is to exemplify how ethical dilemmas emerge from what might at first sight appear as technical considerations. More specifically, studying ethical choices based on computer simulations, this paper aims to show that even if an algorithm is based on an established ethical principle, in our case the maximization of expected utility over time, implementing such a principle within a decision procedure requires many choices to be made that are actually open to reasonable disagreement, thus bringing us back to Moore's (1959) open question. Starting from the implementation of rational choice theory within ethical decision making through risk assessment as motivated by a recent example within the machine ethics literature, we define a Python class `Risk_Simulation()` as a possible ethical decision procedure in order to exemplify the effect of parameter choice on simulation results and ethical choices. This Python class will allow us to exemplify many technical issues with ethical ramifications that leave room for ethical pluralism, reasonable disagreement and, accordingly, that can be conceived as open questions.

Ethics, rational choice, and computer science

Ethics can be defined as a systematic and rational evaluation of the norms, values and principles that should guide our actions. From this perspective, ethical choices are implicitly (and, at least, partially) taken as rational choices. Despite a lack of consensus over the definition of rationality, *instrumental rationality*, understood as the capacity to take the appropriate means to reach one's ends, is widely admitted as a necessary (though not sufficient) part of rational choice (Broome 2013). When conceived from the perspective of instrumental rationality, rational choice can be characterized as a choice that maximizes expected utility over possible outcomes (Buchak 2013; Jeffrey 1965; Savage 1972). Such an understanding of rational choice has its roots in utilitarianism and, more generally, consequentialism (Sen and Williams 1982). Consequentialism (resp. utilitarianism) promotes the idea that the actions we undertake should be the ones with the highest value (resp. utility). Given that value and utility can be quantified (e.g., pleasure, lack of pain, money, lives saved, etc.), these theories are usually seen as natural candidates for machine ethics (cf. Anderson, Anderson, and Armen 2004). As such, consequentialism, which is a core element of rational choice theory (Verbeek 2008), is considered to be at the very foundation of machine ethics (Moor 1999), thus explaining why the maximization of expected utility is usually presented as a necessary characteristic of rational artificial agents to computer scientists and engineers (e.g., Kochenderfer 2015, Russell and Norvig 2022). To illustrate, following Tolmeijer et al. (2020), around 40% of the listed approaches in machine ethics rely (at least partially) on some form of consequentialism. Cloos (2005), for instance, proposed an Utilitibot based on dynamic Bayesian networks as well as a Markov decision process to allow for autonomous ethical decisions based on the maximization of expected utility.

Beside expected utility, another important aspect of rational choice theory appealing for the automation of decision procedures is risk analysis. When facing a choice, expected utility, defined as a weighted average of all possible outcomes, is computed using the probability (either understood as a personal degree of belief or as a relative frequency; Hansson 1993) and the utility (quantified value) of each possible outcome. When choices are expected to provide undesirable consequences, this probability is interpreted as the risk that such an outcome occurs (Hansson 2004). Thus understood, risk analysis is especially relevant from the perspective of machine ethics. Moor (2006), for instance, considered that an interesting aspect of explicit ethical agents was that they could be "autonomous [in the sense] that [they] could handle real-life situations involving an unpredictable sequence of events". Accordingly, risk analysis is implicitly conceived as a dimension of automated ethical decision procedures insofar as explicit (or full) ethical agents need to be able to properly assess and manage risks in order to properly choose between possible alternatives with uncertain outcomes.

Risk analysis in the long run

In a recent contribution to the machine ethics literature, Thoma (2022) showed that the maximization of expected utility based on risk analysis in the long run leads to what she dubs the *moral proxy problem*. In a nutshell, the moral proxy problem manifests itself through different and incompatible attitudes towards risk depending on whether machines act as proxies for lower-level agents (e.g., technology user) or for higher-level agents (e.g., developers, legislators). While some choices with uncertain outcomes can be presented to lower-level agents as one-time risk analyses (i.e., maximizing expected utility over a one-time choice), the same choice can be understood as occurring multiple times from the perspective of higher-level agents, and legitimate attitudes towards risks (e.g., risk aversion) will vary depending on the agency's level. One of the examples she uses to illustrate this problem is the *Artificial Rescue Coordination Center*. Assume an autonomous algorithm that dispatches emergency vehicles in a context of limited resources and where only one emergency vehicle is available. Her example concentrates on a situation where a choice has to be made between one of two fatal accidents involving respectively one and three individuals. The example is framed as follows: If the vehicle is

dispatched to Accident 1, then one person will be saved for certain (i.e., the probability of succeeding in saving the individual is 1), while if it is dispatched to Accident 2, then there is a .5 probability of saving three persons and a .5 probability of saving none. Using these probabilities and the number of individuals saved as utilities, she thus considers expected utilities of $1(1) = 1$ and $.5(3) + .5(0) = 1.5$ for Accidents 1 and 2, respectively. Understood from the perspective of a lower-level agents, she argues that, in this context, it would be reasonable (or understandable) for the agent to be risk averse and to prefer sending the emergency vehicle to Accident 1 instead of Accident 2. But when the choice is understood from the perspective of a higher-level agent, for instance if the algorithm is to perform a choice between Accidents 1 and 2, say, one hundred times, then, in the long run, the expected utility of always choosing Accident 1 would be 100 lives saved, whereas always choosing Accident 2 would yield 150 lives saved. But more importantly, Thoma (2022) argues that in such a case it would be unreasonable to be risk averse and not choose Accident 2 insofar as there would be less than a .5% "chance of saving fewer lives than if one always went for Accident 1". On these grounds, she concludes that, from the perspective of a higher-level agent expecting the algorithm to make the choice many times during its life cycle, one should always choose Accident 2. Thus the moral proxy problem.

Decision procedure based on risk analysis

Thoma's (2019, 2022) analysis was taken as a starting point of our investigation given that it builds on consequentialism, rational choice theory, and risk analysis, which are well established within the machine ethics literature and are key elements of autonomous ethical agents (Kochenderfer 2015, Russell and Norvig 2022). Under the assumption that there is a solution to the moral proxy problem in favor of higher-level agents, we defined an algorithm focusing on risk analysis to justify the choice that should be made between Accidents 1 and 2 from the perspective of a higher-level user, allowing us to study the effect of parameter change and coding on a hypothetical machine's decision. Inspired by Thoma's analysis, where risk is conceptualized as the proportion of cases where fewer lives are saved given the *a priori* distribution of possibilities, we investigated how such a decision procedure would fare if an algorithm was defined to choose between Accidents 1 and 2 based on the risk of saving fewer lives in the long run. To accomplish this, we generalized on i) the number x of individuals involved in Accident 2, ii) the probability pr of saving the individuals in Accident 2, and iii) the number n of times the algorithm would face the choice between Accidents 1 and 2. Using Python as well as the library NumPy, we defined a class `Risk_Simulation()` to simulate a series of choices between 1 individual saved for certain (Accident 1) and x individuals saved with fixed probability pr (Accident 2). Rephrasing Thoma's example in order to make explicit the conflicting values underlying the choice to be made between Accident 1 and 2 (i.e., implicit to this example is the idea that the individuals in Accident 2 [resp. 1] will die if the algorithm chooses Accident 1 [resp. 2]), the expected utility of choosing Accident 1 was defined by $1(1 - x)$, whereas the expected utility of choosing Accident 2 was defined by $pr(x - 1) + (1 - pr)(-x - 1)$. Each sequence of length n (representing a scenario where the algorithm would send the emergency vehicle n times to Accident 2) was obtained using the method `single_sequence()` through an iteration of random choices (weighted by their respective probability pr and $1 - pr$) between succeeding and failing to save the individuals in Accident 2. The benchmark of comparison used to determine whether fewer lives would be saved by always choosing Accident 2 was defined by $n(1 - x)$, that is, the expected utility of always choosing Accident 1. Each analysis was based on 10 000 simulations of sequences of length n, representing 10 000 possible outcomes if one were to always choose to try and save the individuals in Accident 2, and where the total utility of each sequence (i.e., number of lives saved within that sequence) was compared to the benchmark value (i.e., utility of always choosing Accident 1) in order to determine the percentage of the 10 000 simulations with total utility below benchmark. Following Thoma, .5% was taken as the cut-off value below which it would be irrational to not always choose Accident 2. Main results were obtained with parameters $pr = [.25, .5, .75]$, $x = [3, 5, 10, 15, 20, 25]$, and $n = [5, 10, 15, 20, 25, 50, 75, 100]$. Given a fixed probability pr, results show that the decision procedure is influenced by the number x of individuals involved in Accident 2 as well as the number n of times the algorithm is expected to face such a choice during its life cycle. As a general rule (see Table 1 for selected results), greater values of n were required for smaller values of x to reach the benchmark, as values of x and n needed to be bigger as the probability pr grew smaller. To illustrate, $pr = .5$ reached the benchmark at $n = 75$ for $x = 3$ as well as $n = 15$ for $x = 5$, whereas with $pr = .25$ the benchmark was reached at $n = 50$ for $x = 10$.

	$x = 3$	$x = 5$	$x = 10$	$x = 15$	$x = 20$
$pr = .5$					
$n = 5$	18.67	2.97	3.03	3.42	3.09
$n = 10$	17.28	1.04	**.13**	.07	**.10**
$n = 15$	6.15	**.36**	.09	.00	.00
$n = 75$	**.10**	.00	.00	.00	.00
$pr = .25$					
$n = 15$	69.15	23.56	8.15	1.28	1.62
$n = 20$	78.59	22.40	2.66	2.31	.26
$n = 25$	85.84	20.83	3.24	.82	.69
$n = 50$	90.22	16.64	**.18**	**.06**	.01

Table 1: Selected Results - Percentage of 10 000 simulations below benchmark for $pr = .5$ and $pr = .25$

Risk assessment and parameter choice

From an ethical standpoint, it should be highlighted that even if we assume that maximizing expected utility is the ethical and rational choice to make over time, there are important aspects that might seem technical at first glance but

that are actually open to reasonable disagreement (cf. Levi 1986). Indeed, results show that the choice prescribed by the decision procedure (i.e., whether one should always choose Accident 2 based on the risk of saving fewer lives) is influenced by the values taken as parameters pr, x and n. As a result, this example can be used to argue that algorithms (at least with respect to automated decision procedures) are not value neutral (cf. Miller 2021): There are technical aspects that can be the object of an ethical evaluation, in this case the choice of parameters to be made by a higher-level user.

Looking back at the decision procedure, a first thing to highlight is that fixing the threshold at .5% of sequence below benchmark as a cut-off value between acceptable and unacceptable risk to justify always sending the emergency vehicle to Accident 2 is, in itself, arguable. On the one hand, whether .5% is indeed an objective value that can discriminate between justifiable and unjustifiable risks is an open question. Some might say that this percentage is so small it would simply be unreasonable to sacrifice an overall greater expected utility for the small probability that we end up with less than the benchmark value (e.g., Thoma 2022). Others might simply see this as an arbitrary value. Why not .51%? On the other hand, individuals with different attitudes towards risk (e.g., risk tolerance or risk aversion) will (reasonably) disagree on whether .5% of sequences with total utility below benchmark is an acceptable risk (cf. Hansson 2003, 2005). As a consequence, there is reasonable disagreement to be expected regarding which threshold value should be chosen, as well as whether this threshold value really discriminates between acceptable and unacceptable risks.

While some might think that fixing the threshold value at 0% would put an end to the debate, it is noteworthy that this would also be open to reasonable disagreement insofar as it would be a false representation of the actual risk. Even if the simulation outputs 0% of sequences with total utility below the benchmark value, this does not mean that it is impossible to obtain such a sequence and that there is no risk. For instance, there is always the possibility of failing to rescue the individuals in Accident 2 at each occurrence of the choice. Though it is true that the percentage of sequences with total utility below the benchmark value tends to decrease as n increases, obtaining 0% of sequences below benchmark only happens because the number of simulations (10 000) is way below the *a priori* distribution for sequences of length $n \geq 14$ (2^{14} = 16 384). As an example, 10 000 simulations only represent .03% of the possibilities for sequences of length $n = 25$. From a technical standpoint, it is understandable to consider a number of total simulations that is below the *a priori* distribution given that there is otherwise an important practical cost (i.e., computation time). From an ethical standpoint, however, one must keep in mind that such a simulation provides us with a fallible approximation of the *a priori* distribution of possible outcomes of sequences of length n. To consider an average of 0% of sequences below benchmark as an acceptable cut-off value would blind us to the fact that there is a real risk of obtaining a series of event with total utility below benchmark (cf. Taleb 2010).

While the number of individuals involved in an accident is objective enough, there is also reasonable disagreement to be expected regarding both the probability that the individuals will be saved and the number of times we expect the algorithm to make such a choice between Accidents 1 and 2. Although it is generally not conceived or presented as such in the computer science literature, it should be emphasized that risk analysis is inherently ethical (Hansson 2003). Indeed, risks cannot be reduced to simple probabilities: In addition to involving values and rights, risks relate to important notions including agency, consent, and equity. Hence, one must be quite careful when conceiving risk analysis as a simple scientific or technical evaluation of an event's probability. When asking the question " what would be *the* correct probability to assign to succeeding in saving the individuals in Accident 2?", the honest answer should be that we simply don't know. As Hansson (2009) argued, we should be wary of choices made "as if reasonably reliable probability estimates were available for all possible outcomes". In this case, we would need more information (e.g., gravity of the injuries, type of accident, meteorological conditions, etc.) regarding each instances of the choice between Accidents 1 and 2 in order to make an informed judgment. Some scholars might see this lack of information that would otherwise provide us with reasons to believe that success is more probable than failure (or vice versa) as an argument in favor of applying the principle of indifference and fixing the probability at .5. The principle of indifference, also known as the principle of insufficient reasons (cf. Keynes 1921; Pettigrew 2014), states that in the absence of reasons to believe in the likelihood of either of these events, success or failure should be considered as equiprobable. Yet, the principle of indifference leads to known paradoxes and has been the subject of various controversies over the years (Dubs 1942; Gilboa 2009; Hájek 2019; Zabell 2016). As a consequence, reasonable disagreement regarding whether or not the principle of indifference should be applied is to be expected. Further, one can also expect different agents to have divergent opinions regarding the probability assessment (cf. Pettigrew 2022) even among those in favor of not applying the principle of indifference. And besides, this would be assuming that *a priori* distributions of possible events or relative frequencies are appropriate ways of assessing risks that individual events occur. Dubs (1942), for instance, (correctly) argued that interpreting the relative frequency of sequences with total utility below the benchmark value to assess the probability that an individual sequence occurs in the real world is an inferential mistake. In light of these considerations, reasonable (and scientific) disagreement is to be expected regarding which value should be chosen as parameter pr.

As for the choice of parameter n, beside the fact that trying to determine how many times the algorithm would face such a choice between Accidents 1 and 2 is at best speculative, there is an interesting point to note regarding the life cycle of the algorithm. Indeed, in the eventuality that such a dispatch algorithm would be put on the market, one could reasonably expect the software to get updates during its life cycle. From the perspective of machine learning, one could even expect the algorithm to learn from data throughout the years. In such contexts, one question that would arise is whether it is indeed the same choice that is iterated by the

machine. Depending on the answer to that question, n can be taken to be either quite large or relatively small. But more importantly, the answer to that question will have a direct impact on the 'right' choice made by the algorithm.

Believe it (or not!)

So far, we have exemplified that there are open questions with ethical ramifications when trying to define a decision procedure based on the maximization of expected utility over time, leaving room for reasonable disagreement as well as a plurality of divergent positions. Now, we wish to show that rational choice theory, here understood as the maximization of expected utility in the long run, is not something that should be seen as outside the realm of ethics (McCarthy 2016). On the contrary, there are deep ethical concerns with respect to the computation of expected utility over time (cf. Hansson 2007). As it happens, the decision procedure provided in the previous section leads to a controversial principle that should (we hope) unsettle even the very profound advocates of rational choice: As a general rule, we obtain that one should always favor Accident 2 even if the probability of failure is very high, as long as there are many individuals involved and the choice is expected to be made often. From Table 1, we can already see that even if one believes there is a 75% chance that the rescue attempt fails (i.e., 25% chance it succeeds), the algorithm should choose to send the emergency vehicle to Accident 2 when there are 10 individuals involved and we expect the choice to be made at least 50 times. If the unsettling character of this example is not convincing enough, consider the four following cases, which trade individuals for the number of iterations of the choice. Accident 2 will be chosen even if the probability that the rescue attempt fails is:

- 90%, if there are 25 individuals and if $n = 125$;
- 95%, if there are 50 individuals and if $n = 300$;
- 99%, if there are 200 individuals and if $n = 2500$;
- 99.5%, if there are 900 individuals and if $n = 1550$.

Depending on one's interpretation of probabilities as degrees of beliefs or as relative frequencies (Hacking 2001), we obtain the following paradoxes (i.e., a proposition that is derivable from the decision procedure but that should not be derivable; cf. Åqvist 2002). Understanding probabilities as degrees of beliefs, we obtain that if there are enough individuals to be saved and the choice is made often enough, it does not matter whether or not one believes they will be saved: Even if one is almost certain that the rescue attempt will fail, one should try anyway. When probabilities are understood as relative frequencies, we get that we should always try to rescue the individuals in Accident 2 even if almost all rescue attempts have failed. While maximizing expected utility is generally considered as a necessary (though not sufficient) basis for rationality (e.g. Bradley 2017; Broome 2013; Jeffrey 1965; Savage 1972), this, we believe, brings to the surface an issue that is intrinsically ethical and that is a puzzle even for rational choice theory. Indeed, theories of decision under risks and uncertainty (cf. Buchak 2013; Gilboa 2009; Hansson 2004; Savage 1972) and, more generally, Bayesianism (cf. Sprenger and Hartmann 2019), advocate that probabilities, either understood as relative frequencies or as degrees of beliefs, are of foremost importance to evaluate the (rational) choices that should be made. Yet, maximizing expected utility over time implies that the probability of succeeding in the rescue attempt is irrelevant. While this goes against the very foundation of rational choice theory, this, from an ethical standpoint, reduces maximization of expected utility to the pitfall of brute aggregation of individuals that have plagued utilitarianism since its inception (cf. Audard 1999; Bentham 1834), where only the total number of individuals matters.

Could this really happen?

Overall, our thought experiment shows that even if beforehand we assume that maximizing expected utility over time is the rational and ethical thing to do, we end up with an open question afterwards: It is arguable whether maximizing expected utility to the extent of neglecting the probability of succeeding in the rescue attempt is indeed an ethical choice. To exemplify that this is not only a purely theoretical thought experiment and that implementing such a decision procedure would have dire repercussions in the real world, consider a case where the algorithm would be sold in 50 countries, which would then use it for 10 years. Suppose that a situation in which a choice between saving one individual for certain (Accident 1) or saving 200 individuals with $pr = 1\%$ (Accident 2) is highly improbable (note that care pile-ups involving more than 150 vehicles do happen in history). Say the probability of such a choice occurring is .01% per year. Over the past 20 years only in Canada, there have been on average 112 679 collisions involving fatalities and/or injuries per year (Government of Canada 2022). Assuming that the probability of having to make a choice between Accidents 1 and 2 is .01%, we get that this choice is expected to occur roughly 11.27 times in one year. Over 50 countries and 10 years, this would yield roughly 5 635 occurrences of the choice, which is well above the 2 500 occurrences required to minimize the risk of not maximizing expected utility. In such a situation, 5 635 individuals that could have been saved for certain would have been sacrificed so that we could try, 5 635 times, to save 200 individuals facing an almost certain death.

Mathem...ethics?

It is interesting to look at this issue from a technical standpoint to illustrate how ethics can emerge from mathematical considerations. The algorithm's decision is based on a simulation of 10 000 sequences, where each sequence is compared to the benchmark value. From a technical standpoint, given fixed parameters, the question one needs to ask is how many times we need the rescue attempt to succeed within a sequence of n attempts in order to obtain a total utility equal to or greater than the benchmark value. Recall that the benchmark is defined by $n(1 - x)$. Let m be the number of times the rescue attempt succeeds, and $n - m$ the number of times it fails. The total utility of one sequence is given by $m(x - 1) + (n - m)(-x - 1)$. As such, our inquiry takes

the form of solving equation 1.

$$n(1-x) = m(x-1) + (n-m)(-x-1) \qquad (1)$$

From equation 1, we get that $n = xm$ and, therefore, that $1/x = m/n$. As such, $1/x * 100$ gives us the percentage (or relative frequency) of succeeding rescue attempts needed for the sequence to have a total utility greater than or equal to the benchmark value. Interpreting this relative frequency as the rescue attempt's probability of success pr, it follows that pr tends to be smaller as x grows larger. Furthermore, this probability is actually a lower bound for the simulation. Indeed, running simulations for $n = [5, 10, 15, 20, 25, 50, 75, 100]$ and $x = [5, 10, 15, 20]$ with $pr = 1/x$, we obtain a percentage of sequences with total utility below the benchmark value that keeps oscillating (roughly) between 30% and 70% without diminishing as it normally would when the sequence grew longer. To exemplify, running 10 000 simulations with $pr = .005$, $x = 200$, and $n = 3500$ provided an average of 50.92% of sequences below the benchmark value, whereas simulations with $pr = .01$, $x = 200$ and $n = 2500$ resulted in .33% of sequences below benchmark. Accordingly, a small increase (in this case, .5%) in probability from the lower bound $1/x$ allowed the percentage of sequences below the benchmark value to reach .5% over time.

From a pragmatic standpoint, it should be stressed out that this relationship between the prior probability of succeeding in the rescue attempt (here understood as a degree of belief) and the number of individuals involved in Accident 2 is a bit counter intuitive. Indeed, the though experiment assumes that there is only one emergency vehicle available, and that the individuals involved in the accidents are in such a critical condition that they will likely die if no rescue is provided. Under these critical circumstances, it seems more likely to succeed in attempting to rescue 3 individuals compared to 200. As such, one would expect that the probability of success needs to be greater (e.g. in light of the circumstances and the specificity of the context) when more individuals are involved in order to justify sending the emergency vehicle to Accident 2. Put differently, at first sight, it is not likely that one emergency vehicle will be able to save 200 individuals in critical states. One therefore needs good reasons to believe it can. It is (arguably) not rational to expect that a unique emergency vehicle will be able to save 200 individuals which, we believe, only have a 1% chance of survival. Yet, the relationship between pr and x goes the other way around. It dictates that the probability of success can be quite lower when 200 individuals are involved. If there are 200 individuals involved, for instance, one only needs to think they have a bit more than a .5% chance of survival in order to maximize expected utility in the long run, whereas one needs to believe that 3 individuals have at least (a bit more than) 33.34% chance of survival. From the perspective of risk management, it could be argued that one could tolerate more risks (i.e., lower probabilities, more critical conditions) when less individuals are involved insofar as the more there are individual involved, the less likely it is the emergency vehicle will be able to save them all, which goes against the relationship between pr and x.

Looking forward

Scholars in the scientific community are advocating that AI should (among other things) be safe, reliable, and ethical (cf. Batarseh, Freeman, and Huang 2021). Our endeavor is to contribute to the understanding of ethical AI. As many argue that communication should be improved and words should be chosen carefully so that people can understand what AI is as well as its scope and limitation (cf. Ryan 2020), we believe the word 'ethics' should be used with care and parsimony in order to avoid confusion regarding what ethical AI is and, perhaps more importantly, what ethical AI is not. To be clear, our point is not to argue that scholars should stop attempting to implement ethical theories within machines, nor to argue that such attempts are meaningless. We do believe it is possible to define algorithms and machines able to make choices based on the formal representation of ethical theories (i.e., explicit ethical agents in Moor's sense). However, we think it is a mistake to present these algorithms as ethical agents making choices that are justified from an ethical standpoint. Implementing ethical theories should be done carefully while keeping in mind the limitations of such attempts. Focusing on artificial ethical agents amounts to anthropomorphise AI (cf. Ryan 2020) and blinds us to what ethical AI really is, presenting the agents (and their choices) as if they should be accepted because their behavior are justified from a normative standpoint. From an argumentative perspective, such a use of 'ethics' amounts to an appeal to authority, as if the choices were indeed 'the' ethical choices to be made, while in fact this question is (and should remain) open.

Ethical AI should always be considered in light of specific technologies and in relation to individuals, for instance bearing in mind the purpose of technologies, how they are used, and how they affect people. Considering autonomous technologies as ethical entities in themselves otherwise results in unsafe and unreliable AI. Our example shows how technical choices are laden by ethical considerations, and how automated decision procedures are biased by things as simple as parameter choice. We strongly encourage scholars to not only recognize ethical pluralism as a limitation for machine ethics, but also to be aware that implementing ethics is an everlasting process that requires making choices that are in themselves open to reasonable disagreement. Ethical AI will only be achieved by being aware of the limitations surrounding the automation of ethical reasoning, not by autonomous ethical agents. Ethical AI is an ideal we should aim to reach, and whether or not it is actually reachable is open to reasonable disagreement.

Acknowledgments

This work was financially supported by the *Fonds de Recherche du Québec* [2023-NP-310505] and was further supported in part by funding from the Social Sciences and Humanities Research Council. Special thanks to Olivier Roy for comments and discussions.

References

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

Anderson, M.; Anderson, S. L.; and Armen, C. 2004. Towards machine ethics. In *AAAI Workshop on Agent Organizations: Theory and Practice*.

Åqvist, L. 2002. Deontic logic. In Gabbay, D. M., and Guenthner, F., eds., *Handbook of Philosophical Logic*, volume 8. Kluwer Academic Publishers, 2nd edition. 147–264.

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

Batarseh, F. A.; Freeman, L.; and Huang, C.-H. 2021. A survey on artificial intelligence assurance. *Journal of Big Data* 8(1):1–30.

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

Bradley, R. 2017. *Decision Theory with a Human Face*. Cambridge University Press.

Broome, J. 2013. *Rationality Through Reasoning*. Wiley Blackwell.

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

Buchak, L. 2013. *Risk and rationality*. Oxford University Press.

Cloos, C. 2005. The utilibot project: An autonomous mobile robot based on utilitarianism. In *AAAI Symposium on Machine Ethics*, 38–45.

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.

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

Dignum, V. 2021. The role and challenges of education for responsible AI. *London Review of Education* 19(1):1–11.

Dubs, H. H. 1942. The principle of insufficient reason. *Philosophy of Science* 9:123–131.

Gilboa, I. 2009. *Theory of Decision Under Uncertainty*. Cambridge University Press.

Government of Canada. 2022. Canadian motor vehicle traffic collision statistics: 2019. https://tc.canada.ca/en/road-transportation/statistics-data/canadian-motor-vehicle-traffic-collision-statistics-2020. Accessed: 17-10-2022.

Hacking, I. 2001. *An Introduction to Probability and Inductive Logic*. Cambridge University Press.

Hansson, S. O. 1993. The false promise of risk analysis. *Ratio* 6(1):16–26.

Hansson, S. O. 2003. Ethical criteria of risk acceptance. *Erkenntnis* 59(3):291–309.

Hansson, S. O. 2004. Weighing risks and benefits. *Topoi* 23(2):145–152.

Hansson, S. O. 2005. Seven myths of risk. *Risk Management* 7(2):7–17.

Hansson, S. O. 2007. Risk and ethics. In Lewens, T., ed., *Risk: Philosophical Perspectives*. Routledge. 21–35.

Hansson, S. O. 2009. From the casino to the jungle. *Synthese* 168(3):423–432.

Hájek, A. 2019. Interpretations of probability. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition.

Jeffrey, R. C. 1965. *The Logic of Decision*. University of Chicago Press.

Johnson, D. G. 2011. Computer systems: Moral entities but not moral agents. In Anderson, M., and Anderson, S. L., eds., *Machine Ethics*. Cambridge University Press. 168–183.

Keynes, J. M. 1921. *A Treatise on Probability*. Macmillan and Company.

Kochenderfer, M. J. 2015. *Decision Making Under Uncertainty: Theory and Application*. MIT Press.

Levi, I. 1986. *Hard Choices: Decision Making Under Unresolved Conflitc*. Cambridge University Press.

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

McCarthy, D. 2016. Probability in ethics. In Hájek, A., and Hitchcock, C., eds., *The Oxford Handbook of Probability and Philosophy*. Oxford University Press. 705–737.

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

Moor, J. H. 1999. Just consequentialism and computing. *Ethics and Information Technology* 1(1):61–65.

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.

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.

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

Pettigrew, P. 2014. Accuracy, risk and the principle of indifference. *Philosophy and Phenomenological Research* 92:35–59.

Pettigrew, R. 2022. Aggregating agents with opinions about different propositions. *Synthese* 200(372).

Russell, S., and Norvig, P. 2022. *Artificial Intelligence: A Modern Approach*. Global Edition, 4th edition.

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

Savage, L. J. 1972. *The Foundations of Statistics*. Dover Publications.

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

Sprenger, J., and Hartmann, S. 2019. *Bayesian Philosophy of Science*. Oxford University Press.

Taleb, N. N. 2010. *The Black Swan: The Impact of the Highly Improbable*. Random House Trade Paperbacks.

Thoma, J. 2019. Risk aversion and the long run. *Ethics* 129:230–253.

Thoma, J. 2022. Risk imposition by artificial agents: The moral proxy problem. In Vöneky, S.; Kellmeyer, P.; Müller, O.; and Burgard, W., eds., *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*. Cambridge University Press. 50–66.

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.

Verbeek, B. 2008. Consequentialism and rational choice: Lessons from the Allais paradox. *Pacific Philosophical Quarterly* 89(1):86–116.

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

Zabell, S. 2016. Symmetry arguments in probability. In Hájek, A., and Hitchcock, C., eds., *The Oxford Handbook of Probability and Philosophy*. Oxford University Press. 315–340.