OPEN ACCESS
International Journal on Artificial Intelligence Tools
Vol. 33, No. 3 (2024) 2460001 (24 pages)
© The Author(s)
DOI: 10.1142/S0218213024600017

World Scientific
www.worldscientific.com

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

Understanding the Limits of Explainable Ethical AI

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

Jan Broersen
Utrecht University, Janskerkhof 13, 3512 BL Utrecht, The Netherlands
j.m.broersen@uu.nl

Received 28 July 2023
Accepted 17 November 2023
Published 9 January 2024

Artificially intelligent systems are nowadays presented as systems that should, among other things, be explainable and ethical. In parallel, both in the popular culture and within the scientific literature, there is a tendency to anthropomorphize Artificial Intelligence (AI) and reify intelligent systems as persons. 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 (explainable) from a normative (ethical) standpoint. Under this assumption, and given that utilities and risks are generally seen as quantifiable, many scholars have seen consequentialism (or 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. While some see this implementation as unproblematic, there are important limitations to such attempts that need to be made explicit so that we can properly understand what artificial autonomous ethical agents are, and what they are not. From the perspective of explainable AI, there are value-laden technical choices made during the implementation of automated ethical decision procedures that cannot be explained as decisions made by the system. Building on a recent example from the machine ethics literature, we use computer simulations to study whether autonomous ethical agents can be considered as explainable AI systems. Using these simulations, we 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

This is an Open Access article published by World Scientific Publishing Company. It is distributed under the terms of the Creative Commons Attribution 4.0 (CC BY) License which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

behavior and ethical AI and, incidentally, to raise awareness on the limits of so-called autonomous ethical agents.

Keywords: Ethics of artificial intelligence; ethical pluralism; autonomous ethical agents; automated behavior; automated reasoning.

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

1. 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,[1] scholars are attempting to define algorithms that would allow machines to perform autonomous ethical choices. Tolmeijer *et al.*[2] 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.[3–5] Yet, despite these concerns and limitations, scholars are pursuing the idea that truly autonomous ethical machines can be defined.[6,7] To exemplify, Anderson and Anderson,[6] building on Moor's[8] 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 benefits 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[8]). To some extent, this position is understandable insofar as autonomous technologies are becoming more and more advanced, with reaction times so quick that it can be hard to keep humans in the loop of the decision procedures. Hence, there are reasons supporting that one might want to try to implement ethics within machines. However, such endeavors can be problematic. In Anderson's and Anderson's view, for instance, "ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas." Accordingly, 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 that would be able to unequivocally solve ethical conundrums.

In reaction to this trend in machine ethics, Peterson and Hamrouni[1] argued that such attempts are misconstrued, for there is no such thing as an ethical choice

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

without (among other things) responsibility, compassion and compromise (see also De Cremer and Kasparov[9]; Dignum[10,4]). Building on Moore's[11] 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 (and should) always be open and subject to reasonable disagreement. Their argument can be rephrased as follows. If there exists a unique solution $f(x)$ to ethical dilemmas, if $f(x)$ can be known, and if $f(x)$ can be implemented, then $f(x)$ cannot be the object of an ethical evaluation, since otherwise the question whether $f(x)$ is indeed ethical would not make sense (i.e. $f(x)$ would be ethical by definition). However, it should always be possible to question whether a machine operated by such a function is indeed ethical. Accordingly, whether $f(x)$ is indeed the unique solution to all ethical dilemmas should remain an open question: One can always legitimately ask whether a function $f(x)$ meant to output ethical solutions to dilemmas is indeed correctly defined, correctly implemented, or whether the output is indeed 'the' right one (i.e. whether it is indeed 'the' unique correct solution to the dilemma). As such, $f(x)$ should always be the object of a possible ethical evaluation, implying that $f(x)$ cannot be sufficient to determine what is ethical. From this, it follows that either a unique solution $f(x)$ to ethical dilemmas does not exist, $f(x)$ cannot be known, or $f(x)$ cannot be implemented.

To be explicit, Peterson's and Hamrouni's analysis relies on ethical pluralism, which can be understood to have (at least) two different meanings. On the one hand, the domain of possible ethical evaluations is plural, that is, there is a plurality of things that can be the object of an ethical evaluation. For instance, one can evaluate norms, actions, consequences, risks, algorithms, or individuals to assess whether they are good (bad), (un)fair, (un)just, (un)ethical, (un)acceptable, (un)justified, etc. On the other hand, there is pluralism at the level of normative theories insofar as there exists many ethical positions (e.g. deontology, consequentialism, virtue ethics, etc.) focusing on specific aspects of ethical dilemmas (e.g., deontology focuses on intentions and actions, consequentialism on the value of outcomes, and virtue ethics on individuals and character traits) and that can reasonably be defended from an ethical standpoint. By endorsing ethical pluralism, one thus acknowledges incompatible but complementary perspectives on ethical dilemmas. Ethical pluralism thus 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 (i.e. no aspect is intrinsically superior to others; see also Maclure[12] and Weinstock[13]). Thus understood, ethical pluralism can be seen as an adequate (antirealist) answer to moral relativism allowing one to explain why some actions or choices are excusable while others are not. This element is important insofar as, from a foundational (and epistemic) antirealist standpoint, if there exists a true ethical theory, then this theory cannot be known with certainty,

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

for empirical and scientific knowledge is by definition uncertain and at best probable.[14–16] Accordingly, epistemic disagreement on an alleged unique true ethical theory is doomed to prevail. From the perspective of ethical pluralism, there is no such thing as 'the' correct ethical choice (on this point, and in contradistinction to ethical pluralism, see also Dancy[17]): There is only a space of (mutually exclusive) excusable or understandable choices that are available given specific circumstances. And this is precisely why ethical choice implies (among other things) responsibility, because an 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 upshot of Peterson's and Hamrouni's analysis is to highlight the fact that, by acknowledging ethical pluralism as a limit to ethical AI, autonomous ethical agents will always fail to provide us with unequivocal ethical choices or behaviors. Accordingly, it is a mistake to present autonomous ethical agents as machines and algorithms that could behave "exactly how [they] ought to behave in ethical dilemmas". The aim of the present paper is therefore 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 can 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 value-laden and actually open to reasonable disagreement, thus bringing us back to Moore's 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 explain how autonomous ethical decision making would occur in such conditions and 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.

2. Ethics and XAI

It is now well established that AI systems need to be (among other things) explainable and ethical. While the scientific literature on XAI keeps growing,[18–22] scholars are beginning to realize that, in spite of all the well established ethical principles that were put forward in favor of ethical AI such as trust, fairness, responsibility,

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

well-being, privacy, autonomy, and so on (e.g., see the Montreal Declaration for a Responsible Development of Artificial Intelligence), how to concretely implement these principles is not trivial. As such, although theoretical considerations regarding the principles, values and norms that should guide technological developments are important, there is a pressing need to stop reflecting upon the theory and think of feasible ways to apply this theory to concrete cases. This can be understood as a shift of perspective from normative ethics to applied ethics, emphasizing the importance of understanding how ethical AI can concretely be achieved and, incidentally, the limits of autonomous ethical agents.

The need to ensure that AI systems are ethical has been recognized by the engineering and computer science community. Batarseh *et al.*,[23] for instance, proposed to define AI assurance as "a process that is applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, [...], trustworthy and explainable [...], ethical [...], unbiased [...], and fair". As such, they identified six key features that should be used in the evaluation of AI systems: These systems should be ethical, safe, explainable, fair, secure, and trustworthy. In the scientific literature, there are several surveys and comprehensive reviews on XAI.[18–22] Phillips *et al.*[24] proposed to understand XAI through the satisfaction of four principles, namely that AI systems should be (i) explainable, (ii) meaningful (i.e. understandable by different users and communities), (iii) accurate (i.e. the explanation must be empirically adequate with regard to the system), and (iv) their domain of application should be precisely defined (i.e. cases where the system is not designed to work should be made explicit). One consequence of Peterson's and Hamrouni's analysis is precisely that presenting autonomous ethical agents as something that could unequivocally solve ethical dilemmas violates this fourth principle (i.e. since machines and algorithms cannot unequivocally solve ethical dilemmas, presenting them as if they could is a false representation of their domain of application), thus emphasizing the importance of reflecting on ethical XAI and explainable ethical AI. Before going further, let us review the main notions found within the XAI literature.[21]

Miller[25] argued that the notion of *explanation* in AI could benefit from the understanding we find in the human and social sciences literature. Building on the idea that transparency, interpretability and explainability are prerequisites for trustworthy AI,[26,21] explanations are broadly understood as answers to *why-* (or *how-*, *what-*) questions[27] (e.g. *why* did we obtained the output given the input?[28]). While explanations in science are usually understood on the grounds of causality,[29,30] the assumption that explanations in XAI need to be *causal explanations* is unnecessary. Following Peterson,[15] one only needs a pragmatic account of explanations, which can be understood as empirically adequate[16] answers to why-questions, answers that can as well involve mathematics and statistics.[31,32] As we will see below, this understanding of explanations will allow us to distinguish between different meanings of explainability with respect to autonomous ethical agents as well as bring to light how mathematical considerations are relevant from an ethical standpoint.

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

The second principle advocated by Phillips *et al.*,[24] meaningfulness, refers to understandability, which is sometimes also referred to as interpretability[33] or intelligibility.[21] The interpretability of a system can either be seen to have an intrinsic value (e.g., because otherwise AI can be dehumanizing[26]) or an instrumental value (e.g. to reach trust).[34] Broadly understood, this notion denotes the idea that people with different backgrounds need to be able to grasp how the model works. Understanding how a model works is, in turn, achieved through specific explanations. Given that the depth and the relevance of explanations may vary from one group of agents to another (e.g., developers vs users), explanations should be crafted differently depending on their target audience.[28,25] In the literature, there is also a distinction between local and global interpretability.[19,22] While local interpretability refers to the understanding of how single choices are made, global interpretability refers to the entire model behavior. A similar distinction amounts to distinguishing between post-hoc interpretability, which refers to the explanation of why a particular choice was made, and interpretability as transparency, which rather refers to how the model works.[33,34]

Explainability and interpretability can be understood with respect to the notion of transparent models. Transparent systems are presented as counterparts to opaque (black-box) systems,[18] which usually pertain to machine learning models. Following Arrieta *et al.*,[21] transparent models can be understood in three different senses, namely simulatability (i.e. capacity to be simulated by a human), decomposability (i.e. possibility to explain the model's parts, such as input, parameter, and computation), and algorithmic transparency (i.e. capacity to understand how the model acts given any situation).[21] This latter meaning of transparency will be especially relevant with regard to autonomous ethical agents.

3. 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.[35] When conceived from the perspective of instrumental rationality, rational choice can be characterized as a choice that maximizes expected utility over possible outcomes.[36–38] Such an understanding of rational choice has its roots in utilitarianism and, more generally, consequentialism.[39] 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.[40] As such, consequentialism, which is a core element of rational choice theory,[41] is considered to be at the very foundation of machine ethics,[42]

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

thus explaining why the maximization of expected utility is usually presented as a necessary characteristic of rational artificial agents to computer scientists and engineers.[43,28] To illustrate, following Tolmeijer *et al.*,[2] around 40% of the listed approaches in machine ethics rely (at least partially) on some form of consequentialism. Cloos,[44] 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[45]) 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.[46] Thus understood, risk analysis is especially relevant from the perspective of machine ethics. Moor,[8] 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.

4. Risk Analysis in the Long Run

In a recent contribution to the machine ethics literature, Thoma[47] showed that 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*, which can be seen as a variation of Foot's[48] trolley problem. 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 (see Fig. 1): 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

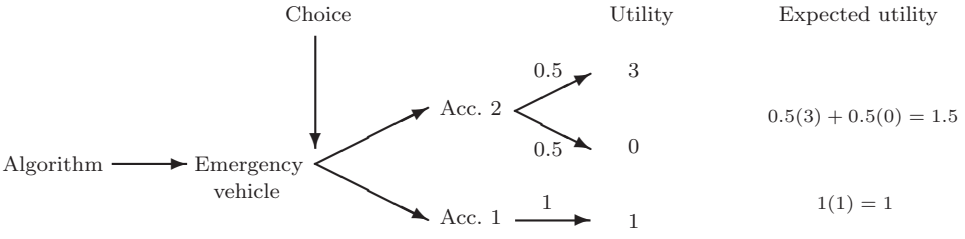

Fig. 1. Artificial rescue coordination center (Thoma).

0.5 probability of saving three persons and a 0.5 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 $0.5(3) + 0.5(0) = 1.5$ 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 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 0.5% "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 (i.e. whether algorithms are considered as proxies for lower- or higher-level agents influences the ethical appraisal of the choices that should be made).

5. Decision Procedure Based on Risk Analysis

Thoma's[49,47] 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.[43,28] Concentrating on 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 (see Fig. 2). 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

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

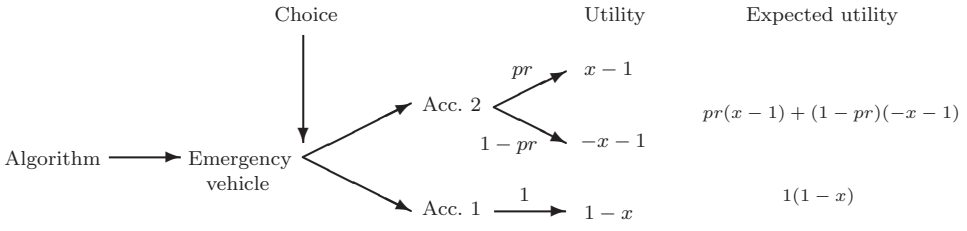

Fig. 2. Artificial rescue coordination center (generalized).

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 Accidents 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, 0.5% 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 $pr = [0.25, 0.5, 0.75]$, $x = [3, 5, 10, 15, 20, 25]$, and $n = [5, 10, 15, 20, 25, 50, 75, 100]$. In terms of global interpretability,[19,22] 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 pattern of behavior (see Fig. 3 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 local interpretability,[19,22] $pr = 0.5$

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

	$x = 3$	$x = 5$	$x = 10$	$x = 15$	$x = 20$
$pr = 0.25$					
$n = 15$	69.15	23.56	8.15	1.28	1.62
$n = 20$	78.59	22.40	2.66	2.31	0.26
$n = 25$	85.84	20.83	3.24	0.82	0.69
$n = 50$	90.22	16.64	**0.18**	**0.06**	**0.01**
$pr = 0.5$					
$n = 5$	18.67	2.97	3.03	3.42	3.09
$n = 10$	17.28	1.04	**0.13**	**0.07**	**0.10**
$n = 15$	6.15	**0.36**	0.09	0.00	0.00
$n = 75$	**0.10**	0.00	0.00	0.00	0.00
$pr = 0.75$					
$n = 5$	1.51	**0.13**	**0.09**	**0.08**	**0.07**
$n = 10$	**0.32**	0.00	0.00	0.00	0.00
$n = 15$	0.03	0.00	0.00	0.00	0.00

Fig. 3. Selected results — Percentage of 10 000 simulations below benchmark for $pr = 0.25$, $pr = 0.5$, and $pr = 0.75$.

reached the benchmark at $n = 75$ for $x = 3$ as well as $n = 15$ for $x = 5$, whereas with $pr = 0.25$ the benchmark was reached at $n = 50$ for $x = 10$.

This model can be conceived to be transparent in all of the aforementioned three senses.[21] Computations can be simulated by a human (although this would be time consuming), all parameters are well-defined and the logic behind the choice is explicit (decomposability), and the model's behavior can be studied given any values for the parameters pr, x and n.

6. 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.[50] 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[51]: There are technical aspects that are value-laden and that can be the object of an ethical evaluation, in this case the choice of parameters to be made by a higher-level user. Notwithstanding the ethical theory that is implemented,

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

programmers and developers will need to make technical choices that are not ethically neutral and that will predetermine the output of the decision procedure.

Looking back at the decision procedure, a first thing to highlight is that fixing the threshold at 0.5% 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 0.5% 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). Others might simply see this as an arbitrary value. Why not 0.51%? Conventions can be questioned from an ethical standpoint. On the other hand, individuals with different attitudes towards risk (e.g., risk tolerance or risk aversion) will (reasonably) disagree on whether 0.5% of sequences with total utility below benchmark is an acceptable risk.[52,53] 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. This is especially relevant from the perspective of XAI if one assumes that autonomous ethical agents can unequivocally solve ethical dilemmas. While the model is explainable in the sense that we can understand why the choice in favor of Accident 1 or 2 was made given the parameters taken as inputs, the model is not explainable with respect to why this choice would be 'the' ethical choice to be made. In explaining why the choice in favor of, say, Accident 2 was made, one would in the end need to refer to the threshold of 0.5% that allows the algorithm to discriminate between acceptable and unacceptable risks. However, considering that this threshold is open to reasonable disagreement, it cannot serve as an explanation of why the choice was (allegedly) 'the' ethical choice to be made by the algorithm.

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 0.03% 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 it is otherwise an intractable problem[28] and there is an important practical cost (i.e., computation time). From an ethical standpoint, however, one must keep in mind that such a simulation provides us

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

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.[54]

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.[53] 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 do not know. As Hansson[55] 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 0.5. The principle of indifference, also known as the principle of insufficient reasons,[56,57] 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.[58–61] 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[62] 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,[58] 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. Again, although appealing to pr can produce a local explanation for the algorithm's output given other parameters, the fact that pr's value is open to reasonable disagreement makes it impossible to appeal to it for an explanation as to why the choice was indeed 'the' correct choice to be made.

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

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.

7. 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.[63] On the contrary, there are deep ethical concerns with respect to the computation of expected utility over time.[64] As it happens, when understood globally, 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 pattern of behavior, 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 (cf. Parfit's[65] repugnant conclusion). From Table 3, 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,[66] we obtain the following paradoxes (i.e., propositions that are derivable from the decision procedure but that should not be derivable[67]). Understanding probabilities as degrees of beliefs, we obtain that if there are enough

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

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. Instrumental rationality in the long run would therefore prescribe that one should try to achieve a goal one does not believe one has the means to achieve. 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. That is, instrumental rationality in the long run would prescribe that we should try to accomplish an action even if all the empirical evidence points to the impossibility of accomplishing such an action.

While maximizing expected utility is generally considered as a necessary (though not sufficient) basis for rationality,[68,35,37,38] 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[36,59,46,38] and, more generally, Bayesianism,[69] 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,[70,71] where only the total number of individuals matters.[65]

8. 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 0.01% 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.[72] Assuming that the probability of having to make a choice between Accidents 1 and 2 is 0.01%, 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 5635 occurrences of the choice, which is well above the 2500 occurrences required to minimize the risk of not maximizing expected utility. In such a situation, 5635

individuals that could have been saved for certain would have been sacrificed so that we could try, 5635 times, to save 200 individuals facing an almost certain death.

9. 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 Eq. (1).

$$n(1-x) = m(x-1) + (n-m)(-x-1) \tag{1}$$

From Eq. (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 $n = [5, 10, 15, 20, 25, 50, 75, 100, 2500]$ and $x = [3, 5, 10, 15, 20]$ 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 converging as it normally would as the sequence grew longer (see Fig. 4). To exemplify,

	$x=3$	$x=5$	$x=10$	$x=15$	$x=20$
$n=5$	46.64	32.62	58.70	70.61	76.79
$n=10$	55.51	37.53	35.01	50.49	59.29
$n=15$	40.11	39.54	54.74	36.26	46.60
$n=20$	47.74	41.37	39.19	61.59	35.05
$n=25$	53.68	41.98	53.71	48.91	64.67
$n=50$	48.68	44.39	42.59	56.95	54.56
$n=75$	46.35	45.38	51.45	44.19	48.12
$n=100$	51.60	44.47	45.06	50.78	43.52
$n=2500$	50.38	49.11	49.02	49.65	48.44

Fig. 4. Percentage of 10 000 simulations below benchmark for $pr = 1/x$.

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

	$x = 3$	$x = 5$	$x = 10$	$x = 15$	$x = 20$
$n = 5$	36.98	24.32	43.61	52.99	59.19
$n = 10$	41.73	23.81	20.27	28.04	34.98
$n = 15$	25.03	23.56	31.74	15.55	20.13
$n = 20$	30.01	22.40	17.21	30.37	12.53
$n = 25$	32.87	20.83	25.48	19.54	27.20
$n = 50$	22.31	16.64	10.77	15.15	11.27
$n = 75$	16.66	12.57	10.64	4.98	5.24
$n = 100$	16.62	10.02	5.56	4.43	2.32
$n = 150$	8.96	6.41	3.00	1.54	1.41
$n = 200$	6.55	3.99	1.37	1.03	**0.40**
$n = 250$	5.30	2.67	0.68	**0.46**	0.30
$n = 300$	3.32	1.84	**0.41**	0.16	0.08
$n = 400$	1.77	0.71	0.15	0.04	0.02
$n = 500$	1.07	**0.41**	0.02	0.00	0.00

Fig. 5. Percentage of 10 000 simulations below benchmark for $pr = 1/x + 0.05$.

running 10 000 simulations with $pr = 0.005$, $x = 200$, and $n = 3500$ provided an average of 50.92% of sequences below the benchmark value, whereas simulations with $pr = 0.01$, $x = 200$ and $n = 2500$ resulted in 0.33% of sequences below benchmark. Accordingly, a small increase (in this case, 0.5%) in probability from the lower bound $1/x$ allowed the percentage of sequences below the benchmark value to reach 0.5% over time (see Fig. 5).

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 thought 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

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

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 0.5% 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.

10. Explainable Ethical AI, or Ethical XAI?

In Sec. 6, we exposed how technical choices (e.g. selecting values for pr, n as well as the threshold to justify sending the emergency vehicle to Accident 2) are open to reasonable disagreement. Given that these parameters predetermine the choice made by the algorithm (i.e. the output), the following problem arises as an inherent limitation to machine ethics: Autonomous ethical decision procedures are biased by value-laden technical considerations, including parameter choice. As it happens, this problem bears consequences on the possible understanding of autonomous ethical agents as XAI systems. To see this, consider how the notion of explanation can have different meanings when analyzing autonomous ethical agents. In Sec. 5, we showed how our model can be considered as explainable, understandable and transparent, for instance allowing us to explain the choice to send the emergency vehicle to Accident 2 by appealing to specific values of x, pr and n, as well as appealing to Eq. (1) to explain the model's behavior (e.g., why the model converges on choosing Accident 2 as x grows larger). However, this explanation does not allow us to explain why this choice would be the right one to be made, that is, it does not allow us to understand why this choice would be justified from an ethical standpoint. This understanding of an explanation as a justification is common within the XAI literature. Indeed, XAI is often considered as an important characteristic of intelligent systems insofar as it can be used to bring to light why agents are 'justified' to behave in the way they do,[19,73,25,74] which in turn can be used to favor trust in AI systems. Thus understood, a justification is thought to explain why "a decision is a good one"[73,25] and, as such, is presented as a proxy for trust: Appropriate explanations allow users and developers to understand why AI systems can provide us with good decisions. In this sense, a justification is understood as an ethical justification, that is, as an explanation providing reasons supporting the idea that a decision is good (i.e. it provides normative reasons; cf. Raz[75]). However, such an understanding of an explanation as a justification can be misleading. Justification

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

in XAI should not be understood as an ethical justification (i.e. providing reasons supporting the idea that the decision is ethical) but should rather be understood as propositional justification[76] (i.e. reasons supporting why the decision occurred). Otherwise, this would result in an inconsistency for explainable autonomous ethical agents insofar as whether or not one adopts ethical pluralism, there can be no ethical explanations of artificial autonomous agents' behaviors. To see this, consider the following reasoning.

When applied to autonomous ethical agents, it might be tempting to interpret XAI systems as procedures that could provide us with 'the' ethical choice (good decision) to be made. However, the notion of an ethical justification (understood as the reasons supporting why a decision is the correct one), which could be used to unequivocally solve ethical dilemmas, brings us to a deeper problem that can be seen from two complementary angles. First, from a broader perspective, the idea of an ethical justification (providing us with 'the' ethical choice to be made) is inconsistent with ethical pluralism. On the one hand, if one adopts ethical pluralism, then, by definition, there is no such thing as 'the' correct ethical choice: There is only a space of mutually exclusive choices that can be made, depending on the normative principles one adopts. As such, ethical pluralism implies an impossibility result for explainable autonomous ethical agents: One will never be able to explain why a particular decision is indeed 'the' ethical decision to be made. Under the assumption of ethical pluralism, automated ethical agents are not explainable in the ethical sense. On the other hand, rejecting ethical pluralism requires the endorsement of a normative theory as 'the' correct (true?) normative theory. Notwithstanding all the problems and difficulties one would face by endorsing such a position[1] (which has been an open problem in philosophy for the past 2000 years), it happens that one would still face an impossibility result with respect to explainable autonomous ethical agents, which brings us to our second point: Even if one assumes the denial of ethical pluralism and, incidentally, one assumes that some specific ethical principle(s) are the right one(s), one will still face the fact algorithms are biased by (value-laden) technical choices (e.g., parameters predetermine the output of the algorithm). Consequently, even if an algorithm is based on established ethical principles (in our case, the maximization of expected utility over time), there is reasonable disagreement to be expected with respect to technical choices that predetermine the algorithm's output and, *a fortiori*, there is also reasonable disagreement to be expected with regard to the algorithm's output. Put differently, as long as there is reasonable disagreement to be expected with regard to technical (value-laden) choices, then the output should not be considered as 'the' correct one insofar as it will also be open to reasonable disagreement. An ethical justification is therefore also not possible in this case. Explainable ethical AI, understood as the explanation of why (from an ethical standpoint) autonomous ethical agents behave correctly, is therefore impossible.

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

11. Ethical Transparency

Summing up, while an explanation does provide us with propositional justification insofar as it presents sufficient reasons answering a why-question[76] (which is consistent with Biran's and Cotton's[73] understanding of a justification as the rationale behind each step of a decision), these reasons are not sufficient to establish whether one is justified to believe that the algorithm produces 'the' correct ethical output (i.e., whether it is *good* in the ethical sense). Although propositional justification refers to the reasons used to support a choice, these reasons do not give one the right to act in such a way.[1] Accordingly, justification with respect to explanation should not be understood as entitlement. Rather, in light of ethical pluralism, an ethical justification would be better thought of as an excuse, where one can be excused to have made a choice given specific circumstances and while other choices, with their respective reasons (i.e. propositional justification) were available. One should keep in mind, though, that algorithms and machines cannot be excused.

Whether one adopts ethical pluralism or not, AI systems cannot be explainable in the ethical sense. That is, one cannot explain why an output would be 'the' correct one insofar as the question will remain open to reasonable disagreement. This follows from the fact that value-laden technical choices are open to reasonable disagreement. One should therefore not speak of explainable ethical AI systems, in the sense that one could have an ethical explanation of the systems' behavior, but should rather speak of ethical XAI systems. As XAI is meant to capture AI systems for which their domain of application is precisely defined,[24] that is, for which one knows the limitations of the system, including when it is meant to work, and when it is not, ethical XAI should be understood as the class of XAI systems for which their domain of application is precisely defined from an ethical standpoint. To be precise, ethical XAI is not meant to say that XAI systems can be trusted to provide us with ethical decisions, bur rather that they do not violate recognized principles such as fairness, trust or transparency. In a sense, ethical XAI can be seen as an overarching notion that characterizes what AI systems should be. For instance, an untrustworthy system would not be taken as ethical (i.e. *ethical* AI implies *trustworthy* AI), whereas it is not the case that an unethical system would necessarily be taken as untrustworthy (e.g. a system can be trusted to provide an unfair result). One important characteristic of XAI systems is that we can understand when they are not supposed to work. Ethical XAI makes it clear that XAI systems are not explainable in the ethical sense and, as such, are not meant to provide us with ethically justified decisions.

The upshot of our analysis is that autonomous ethical agents are not explainable from an ethical perspective. Only propositional justification is available as an explanation, which clearly defines the boundaries of XAI and autonomous ethical agents. We therefore propose to complement the principles of XAI[21] with a fourth sense of transparency, namely ethical transparency, defined as the capacity to understand

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

why specific choices or actions are not unequivocal solutions to ethical dilemmas, including:

- a description of the normative principles assumed;
- a description of competing normative principles that could be taken as reasonable;
- an analysis of the technical biases predetermining the output;
- and an analysis of alternative competing choices that could be taken as reasonable.

12. Looking Forward

Scholars in the scientific community are advocating that AI should (among other things) be safe, reliable, and ethical.[23] Our endeavor is to contribute to the understanding of ethical and explainable 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,[77] 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, including the fact that in addition to a plurality of reasonable and incompatible normative theories, there are competing decision theoretic approaches[78] to autonomous decision making that would face similar problems where they to be implemented. Focusing on artificial ethical agents amounts to anthropomorphise AI[77] 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. For future research, we intend to study how value-laden technical choices emerge during the implementation of different decision theoretic approaches as well as within algorithms that are driven by different normative theories.

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.

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

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. Special thanks to Olivier Roy for comments and discussions. Part of this work was first published in C. Peterson, Further thoughts on defining $f(x)$ for ethical machines: Ethics, rational choice, and risk analysis, *The International FLAIRS Conference Proceedings*, Vol. 36 (2023).

References

1. C. Peterson and N. Hamrouni, Preliminary thoughts on defining $f(x)$ for ethical machines, in *The International FLAIRS Conference Proceedings*, Vol. 35 (2022).
2. S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen and A. Bernstein, Implementations in machine ethics: A survey, *ACM Computing Surveys* (*CSUR*) **53**(6) (2020) 1–38.
3. M. Brundage, Limitations and risks of machine ethics, *Journal of Experimental & Theoretical Artificial Intelligence* **26**(3) (2014) 355–372.
4. V. Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (Springer, 2019).
5. D. G. Johnson, Computer systems: Moral entities but not moral agents, in *Machine Ethics*, eds. M. Anderson and S. L. Anderson (Cambridge University Press, 2011), pp. 168–183.
6. M. Anderson and S. L. Anderson, Machine ethics: Creating an ethical intelligent agent, *AI Magazine* **28**(4) (2007) 15–26.
7. L. Muehlhauser and L. Helm, The singularity and machine ethics, in *Singularity Hypotheses*, eds. A. Eden, J. Moor, J. Søraker and E. Steinhart, The Frontiers Collection (Springer, 2012), pp. 101–126.
8. J. H. Moor, The nature, importance, and difficulty of machine ethics, *IEEE Intelligent Systems* **21**(4) (2006). 18–21.
9. D. De Cremer and G. Kasparov, The ethical AI paradox: Why better technology needs more and not less human responsibility, *AI and Ethics* **2**(1) (2022) 1–4.
10. V. Dignum, The role and challenges of education for responsible AI, *London Review of Education* **19**(1) (2021) 1–11.
11. G. E. Moore, *Principia Ethica* [1903] (Cambridge University Press, 1959).
12. J. Maclure, Context, intersubjectivism, and value: Humean constructivism revisited, *Dialogue* **59**(3) (2020) 377–401.
13. D. Weinstock, Compromise, pluralism, and deliberation, *Critical Review of International Social and Political Philosophy* **20**(5) (2017) 636–655.
14. F. Houle and C. Peterson, *Hors de Tout Doute Raisonnable: La Méthodologie et L'adéquation Empirique Comme Fondements de L'épistémologie du Droit de la Preuve* (Les Éditions Thémis, 2018).

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

15. C. Peterson, Methodological empiricism and the choice of measurement models in social sciences, *European Journal for Philosophy of Science* **8** (2018) 831–854.
16. B. C. van Fraassen, *The Scientific Image* (Clarendon Press, 1980).
17. J. Dancy, *Ethics Without Principles* (Oxford University Press, 2004).
18. P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold and P. M. Atkinson, Explainable artificial intelligence: An analytical review, *Data Mining and Knowledge Discovery* **11**(5) (2021) 1–13.
19. A. Adadi and M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), *IEEE Access* **6** (2018) 52138–52160.
20. D. Minh, H. X. Wang, Y. F. Li and T. N. Nguyen, Explainable artificial intelligence: A comprehensive review, *Artificial Intelligence Review* **55** (2022) 1–66.
21. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila and F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion* **58** (2020) 82–115.
22. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, A survey of methods for explaining black box models, *ACM Computing Surveys* (*CSUR*) **51**(5) (2018) 1–42.
23. F. A. Batarseh, L. Freeman and C.-H. Huang, A survey on artificial intelligence assurance, *Journal of Big Data* **8**(1) (2021) 1–30.
24. P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski and M. A. Przybocki, Four principles of explainable artificial intelligence, in *National Institute of Standards and Technology* (U.S. Department of Commerce, 2021)
25. T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* **267** (2019) 1–38.
26. N. Colaner, Is explainable artificial intelligence intrinsically valuable?, *AI & Society* **37** (2022) 1–8.
27. R. Confalonieri, L. Coba, B. Wagner and T. R. Besold, A historical perspective of explainable artificial intelligence, *Data Mining and Knowledge Discovery* **11**(1) (2021) 1–21.
28. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th edn. (Global Edition, 2022).
29. J. Woodward, *Making Things Happen*: *A Theory of Causal Explanation* (Oxford University Press, 2003).
30. M. Strevens, *Depth*: *An Account of Scientific Explanation* (Harvard University Press, 2008).
31. A. Baker, Are there genuine mathematical explanations of physical phenomena?, *Mind* **114**(454) (2005) 223–238.
32. M. Lange, In defense of really statistical explanations, *Synthese* **200**(5) (2022) 388.
33. Z. C. Lipton, The mythos of model interpretability, *Queue* **16**(3) (2018) 1–27.
34. C. Beisbart and T. Räz, Philosophy of science at sea: Clarifying the interpretability of machine learning, *Philosophy Compass* **17**(6) (2022) 1–11.
35. J. Broome, *Rationality Through Reasoning* (Wiley Blackwell, 2013).
36. L. Buchak, *Risk and Rationality* (Oxford University Press, 2013).
37. R. C. Jeffrey, *The Logic of Decision* (University of Chicago Press, 1965).
38. L. J. Savage, *The Foundations of Statistics* (Dover Publications, 1972).
39. A. Sen and B. Williams, *Utilitarianism and Beyond* (Cambridge University Press, 1982).
40. M. Anderson, S. L. Anderson and C. Armen, Towards machine ethics, in *AAAI Workshop on Agent Organizations*: *Theory and Practice* (2004).

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.

41. B. Verbeek, Consequentialism and rational choice: Lessons from the Allais paradox, *Pacific Philosophical Quarterly* **89**(1) (2008) 86–116.
42. J. H. Moor, Just consequentialism and computing, *Ethics and Information Technology* **1**(1) (1999) 61–65.
43. M. J. Kochenderfer, *Decision Making Under Uncertainty: Theory and Application* (MIT Press, 2015).
44. C. Cloos, The utilibot project: An autonomous mobile robot based on utilitarianism, in *AAAI Symp. on Machine Ethics* (2005), pp. 38–45.
45. S. O. Hansson, The false promise of risk analysis, *Ratio* **6**(1) (1993) 16–26.
46. S. O. Hansson, Weighing risks and benefits, *Topoi* **23**(2) (2004) 145–152.
47. J. Thoma, Risk imposition by artificial agents: The moral proxy problem, in *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*, eds. S. Vöneky, P. Kellmeyer, O. Müller and W. Burgard (Cambridge University Press, 2022), pp. 50–66.
48. P. Foot, The problem of abortion and the doctrine of double effect, *Oxford Review* **5** (1967) 5–15.
49. J. Thoma, Risk aversion and the long run, *Ethics* **129** (2019) 230–253.
50. I. Levi, *Hard Choices: Decision Making Under Unresolved Conflict* (Cambridge University Press, 1986).
51. B. Miller, Is technology value-neutral?, *Science, Technology, & Human Values* **46** (2021) 53–80.
52. S. O. Hansson, Seven myths of risk, *Risk Management* **7**(2) (2005) 7–17.
53. S. O. Hansson, Ethical criteria of risk acceptance, *Erkenntnis* **59**(3) (2003) 291–309.
54. N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable* (Random House Trade Paperbacks, 2010).
55. S. O. Hansson, From the casino to the jungle, *Synthese* **168**(3) (2009) 423–432.
56. J. M. Keynes, *A Treatise on Probability* (Macmillan and Company, 1921).
57. P. Pettigrew, Accuracy, risk and the principle of indifference, *Philosophy and Phenomenological Research* **92** (2014) 35–59.
58. H. H. Dubs, The principle of insufficient reason, *Philosophy of Science* **9** (1942) 123–131.
59. I. Gilboa, *Theory of Decision Under Uncertainty* (Cambridge University Press, 2009).
60. A. Hájek, Interpretations of probability, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2019), Fall 2019 edn.
61. S. Zabell, Symmetry arguments in probability, in *The Oxford Handbook of Probability and Philosophy*, eds. A. Hájek and C. Hitchcock (Oxford University Press, 2016), pp. 315–340.
62. R. Pettigrew, Aggregating agents with opinions about different propositions, *Synthese* **200**(372) (2022).
63. D. McCarthy, Probability in Ethics, in *The Oxford Handbook of Probability and Philosophy*, eds. A. Hájek and C. Hitchcock (Oxford University Press, 2016), pp. 705–737.
64. S. O. Hansson, Risk and ethics, in *Risk: Philosophical Perspectives*, ed. T. Lewens (Routledge, 2007), pp. 21–35.
65. D. Parfit, *Reasons and Persons* (Oxford University Press, 1984).
66. I. Hacking, *An Introduction to Probability and Inductive Logic* (Cambridge University Press, 2001).
67. L. Åqvist, Deontic logic, in *Handbook of Philosophical Logic*, eds. D. M. Gabbay and F. Guenthner, Vol. 8 (Kluwer Academic Publishers, 2002), pp. 147–264, 2nd edn.
68. R. Bradley, *Decision Theory with a Human Face* (Cambridge University Press, 2017).

69. J. Sprenger and S. Hartmann, *Bayesian Philosophy of Science* (Oxford University Press, 2019).
70. C. Audard, *Anthologie Historique et Critique de L'utilitarisme* (Presses Universitaires de France, 1999).
71. J. Bentham, *Déontologie ou Science de la Morale* (Les Classiques des Sciences Sociales, 1834).
72. Canadian Motor Vehicle Traffic Collision Statistics: 2019 (Government of Canada, 2022), https://tc.canada.ca/en/road-transportation/statistics-data/canadian-motor-vehicle-traffic-collision-statistics-2020,
73. O. Biran and C. Cotton, Explanation and justification in machine learning: A survey, in *IJCAI-17 Workshop on Explainable AI* (*XAI*), Vol. 8 (2017), pp. 8–13.
74. R. Warner and R. H. Sloan, Making artificial intelligence transparent: Fairness and the problem of proxy variables, *Criminal Justice Ethics* **40**(1) (2021) 23–39.
75. J. Raz, *Practical Reason and Norms* (Oxford University Press, 1999).
76. J. Turri, On the relationship between propositional and doxastic justification, *Philosophy and Phenomenological Research* **80**(2) (2010) 312–326.
77. M. Ryan, In AI we trust: Ethics, artificial intelligence, and reliability, *Science and Engineering Ethics* **26** (2020) 2749–2767.
78. K. Steele and H. O. Stefánsson, Decision theory, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2020)

Int. J. Artif. Intell. Tools 2024.33. Downloaded from www.worldscientific.com
by UTRECHT UNIVERSITY on 09/26/24. Re-use and distribution is strictly not permitted, except for Open Access articles.