AI & SOCIETY (2024) 39:1891–1903
https://doi.org/10.1007/s00146-023-01640-1

OPEN FORUM

When something goes wrong: Who is responsible for errors in ML decision-making?

Andrea Berber[1] · **Sanja Srećković**[1]

Received: 24 January 2022 / Accepted: 23 January 2023 / Published online: 13 February 2023

Abstract
Because of its practical advantages, machine learning (ML) is increasingly used for decision-making in numerous sectors. This paper demonstrates that the integral characteristics of ML, such as semi-autonomy, complexity, and non-deterministic modeling have important ethical implications. In particular, these characteristics lead to a lack of insight and lack of comprehensibility, and ultimately to the loss of human control over decision-making. Errors, which are bound to occur in any decision-making process, may lead to great harm and human rights violations. It is important to have a principled way of assigning responsibility for such errors. The integral characteristics of ML, however, pose serious difficulties in defining responsibility and regulating ML decision-making. First, we elaborate on these characteristics and their epistemic and ethical implications. We then analyze possible general strategies for resolving the assignment of moral responsibility and show that, due to the specific way in which ML functions, each potential solution is problematic, whether we assign responsibility to humans, machines, or using hybrid models. Then, we shift focus on an alternative approach that bypasses moral responsibility and attempts to define legal liability independently through solutions such as informed consent and the no-fault compensation system. Both of these solutions prove unsatisfactory because they leave too much room for potential abuses of ML decision-making. We conclude that both ethical and legal solutions are fraught with serious difficulties. These difficulties prompt us to re-weigh the costs and benefits of using ML for high-stake decisions.

Keywords Machine learning · Algorithmic decision-making · Opacity · Responsibility · Liability · Hybrid responsibility · Machine responsibility

1 Introduction

We aim to address the core of the ethical debates surrounding the use of decision-making systems based on machine-learning (hereafter: ML systems) for high-stakes decisions.[1] There is an abundant, and increasingly growing literature focused on the ethical problems of using ML systems for decision-making, e.g. in medicine and healthcare, criminal justice, job application assessment, insurance and loan qualification, etc. (Wexler 2017; Rudin 2019; Varshney and Alemzadeh 2017; Flores et al. 2016; Wang et al. 2019; Yeung 2019; Russell and Norvig 2016). The literature is mostly focused on the requests for explanations of the functioning of ML systems or of particular decisions (Miller 2017; Gilpin et al. 2018; Guidotti et al. 2018; Mittelstadt et al. 2019; Paez 2019; Samek et al. 2019; Rudin 2019). The pressure from both the academic literature and the various governmental institutions and NGOs to provide such explanations resulted in the project of Explainable AI (UNI Global Union 2018; Floridi et al. 2018; Hoffman et al.

[1] We take artificial neural networks (ANNs) as the paradigmatic type of ML model. Some aspects of our analysis may, however, be relevant for other types of models that give rise to similar issues.

✉ Sanja Srećković
sanja.sreckovic87@gmail.com

Andrea Berber
berberandrea@gmail.com

[1] Faculty of Philosophy, University of Belgrade, Belgrade, Serbia

2018; Gilpin et al. 2018; Ribera and Lapedriza 2019). It is strongly advocated that a right to an explanation is in itself an important right for the subjects of the decisions (Wachter et al. 2016, 2018; Samek et al. 2019; Goodman and Flaxman 2017; Zednik 2019; Guidotti et al. 2018; Zerilli et al. 2019). Designing various technical tools that provide insight into the functioning of ML is seen as a solution that should satisfy the right to an explanation.

We argue, however, that explanatory efforts only provide a limited value for achieving ethical AI. In cases when an algorithmic decision has already led to detrimental effects, e.g., a mistaken diagnosis, an unfair decision concerning parole release, or even more radically, a large-scale destruction caused by autonomous systems in warfare (Sparrow 2007; Asaro 2012; Apps 2021), it is important to be able to assign responsibility and legal liability, just as in human decision-making. The explanatory efforts concerning ML, however, do not automatically point to the locus of responsibility for the detrimental decisions. Moreover, it seems that there are deeply ingrained reasons why it may be extremely difficult to assign responsibility in the context of ML decision-making. The problems seem to stem from the characteristic way in which ML systems work, and it is not clear whether it is possible to overcome the ethical problems while keeping the advantages that ML provides. This paper aims to explain in what ways the inherent characteristics of ML create obstacles for assigning moral responsibility and legal liability.

The paper is structured as follows. In Sect. 2 we present the (mostly) inherent characteristics of ML systems. In Sect. 2.1 we briefly expose the main advantages of using ML systems for decision-making, which make them appealing for use in a variety of fields. In Sect. 2.2 we indicate the epistemic consequences of the inherent characteristics of ML. In Sect. 2.3 we address the attempts to remedy these consequences through the xAI project, and indicate why these attempts are not sufficient for achieving ethical AI.

In Sect. 3 we analyze the negative aspects of using ML for decision-making, which involve potentially harmful or rights-infringing decisions, and the problem of determining the locus of responsibility. In 3.1 we discuss the relationship between the characteristics of ML and the problems with control and responsibility. In 3.2 we present step-by-step how these issues arise in the ML decision-making process. We show that the problems with assigning responsibility arise from the inherent characteristics of ML presented in Sect. 2.

In Sect. 4 we consider potential ethical (Sect. 4.1) and legal (Sect. 4.2) approaches to solving this problem, and discuss the difficulties of adopting each of the proposed solutions. We conclude that the inherent characteristics of the way in which ML functions give rise to both the advantages of using ML and the difficulties in assigning moral and legal responsibility. This opens the question of the costs and benefits of using the systems which, although highly efficient, are inherently evasive to human control and difficult to regulate.

2 Inherent characteristics of ML

There are several characteristics integral to the way in which ML functions that we find as having significant ethical implications. We will now briefly describe these characteristics, and then proceed to elaborate on their epistemic and ethical consequences.

The first characteristic—*semi-autonomy*—refers to the self-learning of the ML algorithm. In the training phase, a parameter-updating algorithm learns the model from the patterns in the training data. The result of the training process is a mathematical model, which is a function that maps input features to output features and provides predictions for new inputs based on these patterns (Boge and Grünke 2019; Grossmann et al. 2021). For example, in ANNs, learning consists of refining or adjusting the weights of parameters based on the patterns in the data, where the weights indicate the relevance of a parameter for a particular output—a classification or a prediction. In other words, the higher the weight assigned to the parameter, the more important the parameter is for the given output (Hart 2019). For example, if a high weight is assigned to the parameter of age in the criminal risk assessment, this means that age is an important indicator of whether the defendant will be classified as a low-risk or high-risk category. Since the classifications and predictions such as these are determined by the semi-autonomous process of learning, the engineers do not have direct oversight or control over what the machine learns from the data and what classifications or predictions it will give as output.[2]

The second important characteristic of ML systems is *complexity*. Namely, several aspects of ML systems involve unprecedented quantities, such as data and operations involved in the decision-making processes. First, the massive *datasets* that ML systems learn from, often called 'big data', meaning "rapidly collected, complex data in such unprecedented quantities that terabytes (1012 bytes), petabytes (1015 bytes) or even zettabytes (1021 bytes) of

[2] There are different degrees of human involvement in supervised and unsupervised learning, and this implies different degrees of control over the learning process. Supervised learning is characterized by the use of labeled datasets, which requires human intervention to label the data appropriately. In this way humans 'supervise' machines to learn how to correctly classify data. In contrast, in unsupervised learning the machine discovers the underlying structures of unlabelled datasets on its own. Admittedly, even unsupervised modeling needs human intervention with regard to validating output variables to be able to learn from data. However, there is still a significant degree of (semi-)autonomy present in ML that is relevant for the epistemic consequences we discuss in 2.2.

storage may be required", and the data themselves also often have extremely large dimensionality (Wyber et al. 2015). The size of datasets is one aspect of the overall complexity. The other aspects concern the complexity of the ML *models* and involve the great number of paths the information could travel between input and output through hidden layers. For example, as the number of nodes in a layer, as well as the number of layers, grow in an ANN, this leads to having a number of, say, 10^8 of paths along which information can travel between layers (Boge and Grünke 2019).

Another important characteristic of a large number of ML systems is that they may behave *non-deterministically*. In this context, non-determinism means that, upon each execution, the model may produce different outputs based on the same input. The training algorithms can lead to different behaviors of the model, for example, if the same training data is used in a different order. When tested against the test dataset, the models (that were trained on the same training dataset) show slightly different performances. In contrast, deterministic modeling generates consistent output for the same input in each execution. This kind of modeling is used when the relationship between the variables is determined and there is no randomness or uncertainty. Non-deterministic modeling, on the other hand, is used when the relationship between the variables involved is unknown or uncertain, and it is more suitable for finding approximate solutions relying on the likelihood estimation of the probability of events (Mehta 2022; Ombach 2014).

2.1 Pros: why use ML

These inherent properties of ML bring about several advantages over traditional, human decision-making, such as accuracy, but also efficiency and reliability. Machine learning predictions prove to be more accurate than human predictions in many areas, in the sense that they produce fewer errors in classification or prediction tasks (Goh et al. 2020; Lee 2020). ML systems can also process tremendous amounts of data in a short period of time and provide predictions and recommendations for a large number of cases. This high efficiency is crucial in fields where decisions need to be made in a limited time (e.g. healthcare). Since ML systems can also process much larger amounts of data than humans, their capacities allow taking more parameters into account than human decision-makers. This may enable the ML system to capture more regularities and correlations in the data than humans would, but also the correlations that are not obvious to humans as being relevant.[3]

In virtue of these advantages, namely, efficiency in processing a tremendous amount of data and providing highly accurate predictions, ML was found to be suitable for application in many critical areas. In healthcare, for example, ML is immensely helpful in diagnostics, therapy recommendations, and automation of tasks that would take too much time and effort from doctors and patients. ML is able to perform a thorough analysis and organize large datasets with many data points such as patient files, hospital records, etc. For example, using ML in diagnostics (e.g. the InnerEye project designed to differentiate healthy cells and tumors on 3D radiological images) results in an increase in diagnostic accuracy as well as reaching results much faster, which allows starting the treatments earlier. In therapy recommendation, for example, IBM's Watson Oncology system uses the patient history to suggest multiple potential treatment options, thereby taking into account the personal specifics of each patient, such as drug interactions (Tkachenko 2021). Similarly, in the justice delivery system, the application of ML, due to its efficiency, has the capacity to reduce the pendency of cases and the number of unresolved cases in courts, which in turn affects the efficiency of the judiciary system, and ultimately has an impact on people's access to justice (Pant 2021). In the banking and finance sectors, ML has proved highly efficient in fraud and money laundering detection. Fraudulent behavior is detected by ML algorithms that in real-time can examine an enormous number of data points, transaction parameters, and consumer behavior patterns (Sidelov 2021). The ability for self-learning and the capacity to handle enormous amounts of data thus make ML helpful for performing tasks that are too demanding for humans.

2.2 Epistemic consequences

The characteristics of ML that bring about its advantages also have specific epistemic consequences. ML models are often described as opaque or black boxes since the functioning of the model and the paths leading to a particular decision are not fully epistemically accessible and comprehensible to humans.[4] Opacity is particularly relevant in situations when it is necessary to determine whether there was an error in the decision-making process, and what was its

[3] For example, ML might uncover a correlation between the number of physicians a patient visits, a patient's access to transportation, and a patient's disease outcomes (Russ 2021; see also Wang et al. 2022).

[4] The degree of insight into ML models is of course not the same for an engineer who develops these models and for a person who is a complete layman. The engineer, unlike the layman, knows the general principles of functioning of the ML model and in that sense it can be said that for people developing the models they are 'gray boxes'. However, when we talk about the blackboxness of models, we are referring to the lack of epistemic insight into the aspects of the working of the ML model which applies to experts as well, not just to laymen. Certain aspects of ML models' functioning are not accessible to any human, and this is what we refer to by 'blackboxness' or 'opacity'.

precise cause. Here we briefly explain the different types of epistemic inaccessibility to gain precision and clarity concerning how ethical problems arise.

Semi-autonomy of the ML models results in *the lack of epistemic accessibility*. This form of opacity has not been addressed sufficiently in the literature, however, it was recognized by a number of authors as separate from the opacity that stems from the complexity of the models (Humphreys 2004; Schembera 2017; Boge and Grünke 2019; Srećković et al. 2022). We emphasize it as a distinct form of opacity because of its relevance for the ethical consequences we discuss in Sects. 3 and 4. The lack of epistemic accessibility concerns the lack of an explicit representation of the learned information. This may be described in different terms, e.g. as a "lack of an explicit algorithm linking the initial inputs with the final outputs" (Humphreys 2004, p. 149). As Matthias (2004) puts it, "[c]onnectionist systems lack an explicit representation, and the contained information can only be deduced from their behavior. (…) [W]e cannot:—have a look at the information that is stored inside the network, and, even more importantly;—see what information is not represented inside it" (p. 178–9.). This means, in short, that even the engineers do not have epistemic access to the exact processes of decision-making that lead from the input to the output.

The second reason for opacity is *the lack of comprehensibility*. One way in which ML is not comprehensible concerns the quantities involved. This makes the paths of information in ML models too numerous for any human to be able to 'walk through' in a normal life span (Boge and Grünke 2019; Lipton 2016). The massive leap in the quantities of elements involved in the ML decision-making processes raises them far beyond human cognitive capacities. More importantly, there is no explanation in human-understandable terms for why the ML model made a particular choice. The problem concerns mapping the information that is represented in the neural network into human-understandable information. Burrell (2016) calls this a mismatch between the mathematical procedures in ML and human styles of semantic interpretation, in the sense that what the NN represents is not suitable for human comprehension.[5]

The third epistemic consequence of the way in which ML functions is *the lack of predictability*. The former two consequences concern the fact that the *processes* between the input and the output are either inaccessible or incomprehensible. The lack of predictability, however, concerns the dependence of output upon the input. Since ML modeling is often non-deterministic in the sense that the same input can lead to a different output, this makes the decision-making based on such modeling inconsistent. For example, based on the same parameters relevant to a decision on a parole hearing, an ML system may lead to different predictions, leading subsequently to different decisions for the subject. In other words, the characteristic way of ML modeling leads to epistemic uncertainty and unpredictability regarding its decisions.[6]

The fourth, separate and somewhat artificial type of epistemic inaccessibility is *the block of access* via proprietary software: the corporations and other legal entities are protecting the code or a part of the code of the software they own, and as a result, other parties cannot be adequately informed about the exact process behind the decision-making. Since it is, in principle, always possible to remove this kind of block of access by removing restrictions on the code, it is not inherent and thus does not have the theoretical consequences that the other epistemic dimensions have. However, since it could still affect the ability to find the source of the potential errors or biases involved in the decision-making (discussed in the following sections), these practical consequences make them also worth considering.

To sum up, the lack of accessibility, the lack of comprehensibility, the lack of predictability, and the block of access are different epistemic obstacles posed by the often unavoidable characteristics of ML decision-making. They are generally present in ML decision-making but may become especially problematic in cases when ML decisions cause harm to the subjects of the decisions.

2.3 Explainable AI

Making decisions about people's lives based on opaque decision-making raises ethical questions of whether this decision-making is trustworthy, justified, and ethical (Wexler 2017; Rudin 2019; Varshney and Alemzadeh 2017; Flores et al. 2016; Wang et al. 2019; Yeung 2019; Russell and Norvig 2016). The standard response to this issue is requesting explanations of decisions. 'The Explainable AI (xAI) Project' is the result of these tendencies, and its aim is often stated as increasing the level of insight into ML models by various technical tools.

However, the xAI project is by no means an easy solution and faces its own challenges. First, there are problems of determining what exactly are the explanations that are being sought in the sense that the very notion of explanation

[5] There have, of course, been many attempts to make the information involved in ML decision-making explainable to humans, by constructing other models trained to produce explanations—the Explainable AI (xAI) Project. We discuss the xAI project and its limitations in Sect. 2.3.

[6] For more details on the differences between unpredictability and other epistemic obstacles such as unexplainability and incomprehensibility, see Yampolskiy (2020).

in this context is underspecified.[7] The term 'explanation' in the xAI literature and in the external requests for explainability mainly appears in the cluster of interconnected but also insufficiently specified notions such as comprehensibility, interpretability, transparency, understanding (Lipton 2016), and there is no consensus on what constitutes either of these notions in the context of machine learning, nor is it clear how to measure them (Molnar 2019). The majority of the technical literature seems to rely on an idiosyncratic conception of explanation, exemplified by any tool that offers any kind of insight into the model's working. Many of the explanation methods offered in this field are thus criticized for various reasons, for being incomprehensible, inconclusive, potentially misleading, or not proper explanations (Mittelstadt et al. 2019; Rudin 2019).

There might be additional limitations of the xAI project that result from the characteristics of ML, which functions semi-autonomously, non-deterministically, and involves tremendous complexity on several levels. Some aspects of ML decision-making are inaccessible for representation by any xAI tools, since they "never take durable, observable forms" and exist in memory only temporarily (Ananny and Crawford 2018). Furthermore, the xAI methods cannot in principle faithfully represent the original model's calculations, since they perform different calculations than the original.[8] If they performed the same calculations, they would, in fact, be equal to the original model and would *ipso facto* be too complex to comprehend (Rudin 2019). The xAI methods, therefore, do not reflect exactly how the decisions were reached. Moreover, because the original model is opaque, we are not in a position to determine how much the xAI representation diverges from the original decision-making. Some authors find this problematic for reaching an adequate explanation, and raise the question of the extent to which the xAI itself can be trustworthy, and whether it can grant trustworthiness to the original model (Rudin 2019). It might be responded that every explanation rests on simplification and idealization and as such is not completely faithful to its *explanandum*. The truth is, what constitutes a good explanation is still a matter of debate in epistemology and philosophy of science, and, as we noted at the beginning of this section, is even more unclear in a novel field such as ML.

These are the problems that reflect the current state of the art in xAI and may be remedied with further research. Some authors, however, object to the pursuit of complete transparency of ML decision-making, as it supposedly results from a double standard, since human decision-making is non-transparent as well. Human brain is also considered a black box, and there is no access to all the steps that human decision-makers implement in reaching decisions (Zerilli et al. 2019). A human judge, for example, may make a decision about a parole hearing that is racially biased but also rationalize the decision, a posteriori, by appealing to more acceptable reasons. It may seem that the request for complete transparency of ML decision-making poses unrealistically high standards that are not able to be satisfied in the context of human decision-making. However, there is an important difference between these two contexts. Notwithstanding the non-transparent processes of decision-making, in the human context, the locus of responsibility is clear—the person who has made the decision is responsible if it is established that the decision is unfair and/or has led to harmful consequences. In the context of ML, however, the specificities of its way of functioning pose serious obstacles to determining the locus of responsibility for errors and harm. The xAI methods may indeed reveal some aspects of the functioning of a particular model. However, the attainable level of transparency provided by the xAI may suffice for purposes such as model improvement or providing the end-users with some sort of explanation of the decisions, but it does not suffice for resolving the issue of responsibility. We discuss this issue extensively in Sects. 3 and 4, and show in what ways the unavoidable characteristics of ML pose obstacles to assigning responsibility for errors and harm of ML decision-making.

3 Cons: harm and responsibility

Damaging consequences that may arise from the use of ML decision-making can be divided into separate categories: tangible harm and human rights violations. Although tangible harm and violations of human rights are separate issues, they often coincide. As Yeung (2019) notes, a human rights violation does not have to include tangible harm, and vice versa: (a) when Facebook removed the famous photograph of a naked 9-year-old girl fleeing bombs during the Vietnam War, it could be understood as a violation of the right to freedom of expression and information, even though there was no tangible harm; (b) a self-driving car injuring a wild

[7] For example, the legislative request most commonly cited in the xAI literature is the European Union's General Data Protection Regulation (GDPR), which requests that the subjects of automated decision-making are provided with "meaningful information about the logic involved" in reaching the decision (GDPR, Article 12(2)(f)). The GDPR also states the right of the subjects to "obtain an explanation of the decision reached after such assessment and to challenge the decision" (GDPR, Recital 71). These are the only two mentions of anything related to explanations in this regulation, and these two formulations state two essentially different explanatory requests, one concerning the overall mechanism of the ML model, and the other concerning the path to the model's reaching a particular decision.

[8] For example, explanatory tools commonly consist in surrogate models that solely attempt to capture the input–output trends of the opaque model they are intended to explain, but they employ entirely different features and are thereby not faithful to the original model's computations (Rudin 2019).

animal inflicts obvious tangible harm, but it does not violate a human right (Yeung 2019). Nevertheless, here we consider these consequences jointly since our points apply to both. Since opaque ML models are used for making decisions in many fields that fundamentally affect human lives, such as medical diagnosis and treatment, credit scores in loan or job applications, allowing probation, etc., decisions made in these fields are able to both violate human rights and cause tangible harm to its subjects. For example, the decisions based on a flawed automated process may wrongfully deny someone of their freedom (in criminal justice applications), cause prolonged illness, or even directly harm the subjects of the decisions (e.g. in medical applications). Harmful decisions may be caused either by individual errors or by systematic biases embedded in the ML decision-making processes.

Harmful or rights-infringing decisions which result from faulty decision-making raise the issue of assigning responsibility. Responsibility can be moral or legal. The terms related to legal responsibility, such as accountability and liability are often used interchangeably (Cornock 2011; Yeung 2019). Since these terms are often confused or used inconsistently in the literature, we will use them according to the following definitions. By 'responsibility' we always mean 'moral responsibility', that is, we refer to the abstract ethical concept which, when attributed to an agent, implies that the agent who is responsible for some action or outcome deserves moral praise or moral blame (Zimmerman 1997). Responsibility is sometimes considered closely related to the concepts of 'duty' and 'obligation', which the agent is expected to uphold (Talbert 2022). In the philosophical literature, moral agency is considered a precondition of moral responsibility, and moral agency, in turn, requires properties such as consciousness, intentionality, and free will (Johnson 2006; Sparrow 2007; Asaro 2014 and Hanson 2009).[9] Whenever we are talking about legal responsibility, we will use the term 'liability'. By liability, we refer to the related concept, also attributable to moral agents, which implies that there are effective institutional mechanisms that are expected to be deployed in cases of damaging outcomes, including the mechanisms to appropriately sanction responsible agents.

Although moral and legal responsibilities are most often linked, they are nevertheless independent. We can have cases of moral responsibility in which there is no legal responsibility, as well as cases of legal responsibility without clearly defined moral responsibility. When it comes to regulating an area, such as the use of ML, it would be ideal to determine moral responsibility and build a legal framework based on it. In that case, the agent responsible would also be liable and thus subjectable to the established institutional mechanisms of liability. There are many cases, however, where moral responsibility and legal liability do not coincide. For example, an agent may be found responsible for some harmful outcome, but if there are no institutional mechanisms of liability, the agent would bear only the consequences which remain in the abstract sphere of moral judgment without practical consequences or sanctions.[10] Or conversely, there may be cases where the moral issue is too vague, and it may only be legally regulated by circumventing the issue of moral responsibility and defining liability independently.

Decision-making based on ML systems involves distinct difficulties in assigning both responsibility and liability for damaging decisions. These difficulties arise from the inherent characteristics of ML models themselves and are further aggravated by the complex and dynamic socio-technological context of the development and use of these models. The fact that this type of decision-making involves errors and potential harm is not particularly problematic in itself, since any decision-making system is flawed and prone to harmful errors. However, when it comes to ML decision-making systems, a specific problem arises with attributing responsibility for errors and damage.

In the following sections, we show that the responsibility issues stem from properties that are unavoidable and are an integral part of how ML functions.

3.1 Control and responsibility

In Sect. 2 we discussed how ML decision-making is inherently characterized by semi-autonomy, non-deterministic modeling, and complexity that transcends human cognitive capacities. Here we show how these characteristics reflect on the issues of control and responsibility. First, an ML model autonomously adjusts the decision-making parameters without direct human interference. This means that determining the most crucial factors of decisions is delegated to the model, and humans have no direct control over the processes through which the decision-making proceeds. The lack of control is worsened by the quantitative characteristics of ML decision-making. The volume of information, and the

[9] This understanding of moral agency is what in Moor's terminology characterizes a 'full' moral agent—the only kind of moral agent that we can consider morally responsible. For the complete taxonomy of moral agency, which has become canonical in the literature on this topic, see Moor (2006). We discuss the prerequisites of moral agency in the context of ML in Sect. 4.1.

[10] Certain types of moral transgressions, such as lying in everyday life, are not legally regulated, nor are they expected to be. Giving a false promise to a friend is morally reprehensible, but we will not necessarily end up in court because of it. Of course, some cases of lying such as defamation and false testimony in court are legally regulated, but lying in ordinary daily life usually does not fall into these categories. There is also the possibility that some moral offenses are not legally regulated *yet* because they have only recently emerged, such as those made possible by the development of technology, but are expected to be regulated in the future.

complexity involved in ML decision-making lead to a transcendence of human cognitive capacities (Humphreys 2009; Srećković et al. 2022). The quantitative characteristics thus make it extremely difficult for humans to trace or comprehend the processes of decision-making, which complicates efforts to control the processes.[11]

An additional aspect of ML decision-making that aggravates the problem of assigning responsibility concerns the data that is being fed to the ML model. The data has been gathered over long periods, and only reflects the previous practice of various sectors. The model learns the parameter weights from the correlations among the given data and produces predictions and decisions based on what it has learned. Since algorithms learn from existing patterns, they tend to learn human biases as well. If the data are biased (as it often is), or possesses inaccuracies, this may lead to biased decisions. Quite often, biased models can already be in use (long) before the bias is revealed and fixed. For example, the algorithm widely used by U.S. hospitals and insurers for healthcare allocation was found to prioritize white people over black people when individuals of both groups had equal health conditions. The algorithm based the patients' risk scores on previous medical costs, and ended up identifying as a priority the social groups with higher access to, and ability to pay for medical care (Price 2019). Similarly, the criminological software used for predicting the risk of a future criminal offense in the U.S. judicial system was found to be biased against African American defendants, basing its predictions on the previous practices of allowing and denying parole to defendants with similar background profiles (Angwin et al. 2016). Finally, Amazon's hiring algorithms showed clearly sexist preferences toward male candidates. Since it was programmed to replicate existing hiring practices and the patterns of profiles of those previously employed, "the system taught itself that male candidates were preferable" (Lauret 2019). Even though sometimes biased data is all that is available, we cannot just blame the data and call it a day. This is why engineers look for ways to fix the problem, and the xAI methods can help recognize biases and enable them to take measures to correct them. This is not a straightforward task, and some biases can still slip through despite the measures taken (Zhao et al. 2022). It may be difficult to determine who is responsible for biased decisions while the system, which was developed in good faith and was not known to be biased, was used.[12] Hidden biases in the data can thus additionally blur the locus of responsibility along with the other features we discuss.

In addition to the inherent characteristics of ML, the lack of control is aggravated by external factors. Because of the various advantages provided by these systems in many different fields, there is a growing trend in using ML systems for decision-making. In combination with the number of sectors using ML systems and the number of people subject to automated decisions, along with the fact that these numbers will most probably tend to grow as well (Butler 2016), this would allow for worldwide automated decision-making on a massive scale in a variety of contexts. Even at the current rate of usage, it is not clear whether it is possible for humans to inspect the correctness of the decision-making processes occurring in various fields.[13] With the future expansion of the use of ML systems, the possibility of supervision should be diminished even further.

3.2 Problems at every step of assigning responsibility

A number of steps need to be taken to achieve the adequate assignment of responsibility. Here we describe in more detail how the characteristics of ML complicate this process at each step. The first step of assigning responsibility is to determine whether the ML decision-making process was correct and fair. Due to the characteristics presented in Sect. 2.2, namely, lack of accessibility, predictability and comprehensibility, as well as the external factors such as the massive and growing use of ML, it is not clear how humans could carry out sufficient monitoring of the correctness and fairness. This leads to a general loss of human control in the sense that it becomes exceptionally difficult for humans to

[11] It may be objected that the internal processes of human decision-making are also inaccessible, and perhaps even more complex than ML decision-making. We cannot look into the heads of others (the so-called 'problem of other minds'), so we turn to various social procedures developed for inferring the internal states of other humans (see Matthias 2004). These procedures do not make other minds completely transparent, but might provide some kind of insight about the thought processes, intentions and beliefs of others. Similarly, numerous xAI methods are being developed in attempts to gain insight into ML decision-making processes. So why would this make a problem for ML, but not human decision-making? The key difference is that in the case of human decision-making, the locus of responsibility is clear in most cases. In paradigmatic cases, the person who has made a particular decision is the one who is held responsible for it. In the context of ML, however, there are a number of obstacles that make it highly difficult to find the locus of responsibility for the consequences of the decisions. We discuss these obstacles in detail in Sect. 3.2.

[12] This ambiguity of assigning responsibility would not, of course, apply to cases of intentional biasing of data nor to cases of negligent or reckless use, if, for example, a company did not check the ML system for bias, or continued to use it even after bias is discovered.

[13] Having human experts keep track of the correctness of the ML decision-making may seem as a solution to problems of control and responsibility. However, as Matthias points out, "[w]ere it possible to supply every machine with a controlling human expert, nobody would need the machine in the first place" (2004, p. 177). Besides, it does not seem sensible to employ slower or less reliable systems such as humans to keep track of much more efficient and reliable systems and inspect the correctness of the processes. It would defeat the purpose of using ML decision-making in the first place.

conduct meaningful scrutiny and surveillance over the ML processes necessary for assessing the correctness and fairness of decision-making. Thus a significant (and a growing) number of cases of automated decision-making must remain unexamined, and there are bound to be errors that will never be recognized.

For the remaining cases where inspection did expose errors in the process, the second step—determining the source of error—is obstructed by the epistemic characteristics of ML processes. Due to the lack of accessibility and lack of comprehensibility, it is often difficult to determine whether an error is a consequence of the model itself or of the data on which it was learned. Although there are techniques for determining this indirectly (by making changes and deducing from the behavior of the model), in practice it is not always a realistically manageable task. Additionally, the factor of blocked access due to proprietary software means that third parties (e.g. independent committees or other parties) cannot properly access the causes of errors or biases in the decision-making process. This could practically mean that assignment of responsibility is in the hands of the ML system's private owners, who may not have incentives to investigate the causes of errors or biases.

In the even smaller subset of cases where it was possible to pinpoint the source of error, another difficulty occurs in determining whether the error could have been predicted, as predictability is relevant for assigning responsibility (Yeung 2019).[14] However, the epistemic consequences we have discussed in Sect. 2.2 hinder predictability in multiple ways. Due to the autonomous, non-deterministic, and complex nature of ML processes, or due to the errors coming from the data gathered over long periods, it seems extremely difficult for anyone involved in developing or operating the ML model to predict the errors in the decision-making process. Given that the very technique of machine learning is such that the engineer has no direct and complete control over what decisions the machine will make, it is questionable whether human agents should be held morally responsible for the unpredictable decisions of the machine (see Sect. 4.1 for more details).

Alternatively, even if in some cases it is established that errors and harmful consequences could have realistically been predicted and avoided and that humans are responsible, a complex question arises as to which of all the agents involved in the process should be held responsible. ML models are hardly ever developed and used by a single individual. Rather, the models are commonly developed by a team or even teams of experts, who often compartmentalize the specific development tasks. This problem is named the 'many hands' problem in the literature (Yeung 2019). It may be encountered in fields where the traditional conditions for responsibility, such as intent, knowledge, freedom of action, etc. are distributed over many different individuals, and none of the individuals separately meet all the conditions (Nissenbaum 1996). Since multiple individuals and mostly teams of individuals are involved in making, approving, and using the ML models, the assignment of responsibility becomes a much less straightforward task. There are many candidates: the company that designed the model, the specific people involved in the design or just the people in the management positions in the process, the institution that uses the decision model, or some higher governing body that authorized the use of the model, the government of the state that allowed such automated decision-making, the ML model itself (if machines are afforded the status of moral agents), or a combination of the above candidates.

An additional difficulty in determining the locus of responsibility arises from the socio-technical context of the application of ML models. It is what we call the 'two-way nescience', which arises from the fact that a number of experts who develop ML models often know very little about the fields in which the models will be used (for example, in judiciary or banking systems); and conversely, numerous experts in the fields in which the model is being used know little or nothing about the functioning of the ML (for example, a judge who uses ML system for a parole hearing). This way of usage of ML systems means that the persons who develop them and the persons who apply them are quite often located at different epistemic positions, due to often disparate expertise. Consequently, the errors that may occur in the design or usage of models could result from the difference in the epistemic positions of agents involved (e.g. a developer not properly understanding the details of the judiciary system, or a judge not properly understanding the dependence of the model's score on the data). It is not clear in this context how to correctly assess who should have predicted the detrimental outcomes of the ML model application. In such complex contexts, it is very difficult to achieve an objective assessment of the responsibility of each individual agent because the body assessing liability would have to have expertise in multiple areas, and have insight into all the relevant factors from all areas in which the different agents operate. Making a credible assessment of responsibility may require a dialogue of experts from different sectors involved in a given ML application, which does not appear to be easily achievable in practice.

Finally, for the smallest subset of cases in which all of these obstacles are evaded or overcome, and it is clearly established who are the persons responsible for the harmful consequences, yet another difficulty is posed by the absence of elaborate legal and professional mechanisms that should

[14] Predictability of errors is relevant if we adopt the risk/negligence model of responsibility (Fischer and Ravizza 1998; McKenna 2008; Lunney and Oliphant 2013). For a comparative analysis of different models of responsibility, see Yeung (2019).

ensure that the responsible person really bears the consequences (Mittelstadt 2019).

4 Potential responses to the responsibility problems

At the most general level, there are two main approaches for resolving the issue of responsibility.[15] The first approach would be to resolve the issue of moral responsibility and then to build institutional and legal norms based on that solution. We call this approach 'responsibility first'. Within it, we consider the possible options of attributing moral responsibility for ML decisions, particularly with regard to the fact that both human and autonomous machine actors are involved. In considering this approach, we are primarily interested in potential ways to provide practical guidelines for legally regulating the liability for harmful ML decisions based on moral solutions.[16]

However, since ethical debates are complex and require long-term deliberations, the second, more expeditious approach also seems reasonable to consider. We call this approach 'liability first'. It involves putting the issue of moral responsibility aside and constructing a legal framework that could regulate this area and protect end-users as well as other involved parties without addressing the locus of moral responsibility.

4.1 'Responsibility first'

The first, obvious option is to attribute responsibility to *human* actors (Hall 2001; Goertzel 2002; Johnson 2006; Sullins 2006; Bryson 2010). The rationale behind this option is that the ultimate responsibility for the actions of machines, no matter how intelligent and autonomous they may seem, is always human because it is humans who "determine their [the machines'] goals and behavior, either directly or indirectly through specifying their intelligence, or even more indirectly by specifying how they acquire their own intelligence" (Bryson 2010, p. 65). This is a similar line of thinking to Lady Lovelace's Objection which notes that machines can only do whatever we know how to order them to perform, regardless of the level of their sophistication (Turing 1999; Gunkel 2020). However, as suggested in the previous section, this path is fraught with difficulties. The main difficulty in attributing responsibility to humans is their lack of control over the processes through which ML models reach decisions. The causes for the loss of control—machine semi-autonomy, non-deterministic modeling, and the complexity that transcends human cognitive capacities—open a specific question of the extent to which human beings are morally responsible for the decisions made by machines. Commonly, responsibility is attributed to human beings on the basis of their autonomy and control as decision-makers (Fischer and Ravizza 1998; McKenna 2008; Lunney and Oliphant 2013; Yeung 2019). If the control over the decision-making is transferred to a machine, it may be argued that the responsibility also belongs to the machine and that it would be unjust to hold humans responsible for the machines' decisions (Matthias 2004). Thus, it may seem unjust also to hold humans responsible for the absurdly large number of (semi-autonomously reached and unpredictable) decisions that people have no control over, and over which we cannot realistically expect to be properly monitored or inspected. Attributing responsibility to humans would mean holding human agents responsible for consequences that they could not have foreseen or prevented. Additionally, even if we are prone to accept that humans are responsible, the already mentioned problems of many hands and two-way nescience would pose significant practical obstacles regarding which particular humans we should declare responsible.

Another option would be to hold the *machines* responsible for the decisions they make. The development of autonomous and social machines undermines the pure instrumentalist view of technology, as it seems that autonomous technologies have transcended the role of being mere human tools, and have instead come to occupy the role of human agents themselves (Gunkel 2020). This is most evident in the examples of decision-making systems or self-driving vehicles. The option to assign responsibility to machines may, on the other hand, seem implausible because it is

[15] There is another direction taken in the literature that focuses on building a moral code into the machines. It is considered that this would prevent unethical machine decisions, as well as diminish harm and human rights violations (Anderson and Anderson 2007; Wallach and Allen 2009). However, this project faces several significant challenges. First, it needs to decide on a particular ethical theory: deontological ethics, virtue ethics, utilitarianism, or some other. Second, the chosen theory must be implementable in the machines, in the sense that it must be translatable into a language that allows computation, and it is still unclear whether this is a feasible task. Finally, even if building a moral code into the machines becomes possible, we still need to decide how to deal with errors if they occur. There is no reason to believe that the ethical machines would be completely infallible. It seems that we would still need a principled way of assigning responsibility and dealing with potential errors. It remains for future research to show how successful the project of building moral machines will be in meeting its challenges. Importantly, the topic of this paper is how to assign responsibility for the decisions of ML systems that are currently in use and that do not have any built-in moral code. The discussion might become different if machines with a built-in moral code are used in the future, depending on how exactly they would function.

[16] We will not enter into the controversy over which of the presented directions is the most adequate from the point of view of moral theory in general. We will only briefly present each of the possible directions and analyze the difficulties they face.

generally accepted that attributing responsibility requires moral agency. Attributing moral agency to an entity, in turn, requires the possession of mental qualities such as consciousness, intention, etc. (Johnson 2006; Sparrow 2007; Asaro 2014; Hanson 2009), which does not seem plausible to assign even to the most sophisticated machines, at least at the current stage of development.

Notwithstanding the theoretical debates on the moral status of autonomous machines, there are also important practical problems with this option. Even if we accepted attributing responsibility to machines, it is not clear what practical consequences this would have. What exactly does it mean in practice to consider a machine morally responsible? The question is not only whether machines could be seen as persons and as having genuine moral status; the more important question is whether it would make sense, from a moral (and consequently legal) perspective, to treat machines as persons in the same way that we currently treat human agents and organizations (Gunkel 2020). Thus, one problem with this option is that it has very unclear practical consequences. Another problem is that it is potentially unfair, in that it may lead to underestimating, or even ignoring the further role and responsibility of human agents, and this may be abused for evading the responsibility of humans involved (Johnson and Miller 2008; Gunkel 2020). If the practical consequences are not adequately specified, this direction might boil down to practically holding nobody responsible, which we address in the following passages.

The third option would be to adopt a model of *hybrid* moral responsibility, or 'extended agency theory' that would imply that responsibility is shared between humans and machines (Johnson 2006; Hanson 2009; Verbeek 2011). This option relies on understanding responsibility as distributed across a network of agents involved in the decision-making, including both humans and machines, as well as organizations. It is based on the intuition that all agents involved need to be assigned responsibility for the results of the decision-making. However, this solution also has unclear implications. It still needs to be decided how exactly to distribute the responsibility, namely, what aspects of it belong to machines, and what aspects should be assigned to humans (Gunkel 2020). In addition, it seems extremely difficult to determine who of all the individual agents involved, be it human or artificial, is responsible for which aspect of the harmful decision. Similarly to the option of assigning responsibility solely to machines, this option suffers from having unclear practical consequences, as well as being potentially unjust. Namely, if it is not specified exactly which part or aspect of the responsibility belongs to humans, and which to machines, then the persons who work with the machines, the so-called 'humans in the loop,' might serve as the default culprits whenever the machine reaches a harmful decision. Alternatively, if the shared responsibility was to be interpreted in a contrary way—leaning more to the side of the machines—the presence of a machine in the decision-making could serve as a default excuse for human error and negligence (Siponen 2004; Johnson and Miller 2008; Mowshowitz 2008). Both tendencies of interpreting shared responsibility carry obvious moral dangers. This suggests that it needs to be precisely elaborated on which way the participation of machines in decision-making affects the attribution of responsibility to human agents.

Finally, it could be argued that in some cases *nobody* is responsible for the decisions of the machines. Humans are not responsible because they cannot control machine decisions, provided that the decision-making models are constructed with the best intention. In addition, machines are not responsible because they do not meet the epistemic conditions for being considered moral agents. If the error or bias stems from the data collected over many years, it may seem that there is nobody to whom this type of error or bias should be attributed. Such a scenario makes it implausible to assign responsibility to anyone in particular, even to the machine. The damage caused as a result of machine decisions would then be observed analogously to the consequences of a *force majeure*, such as the damage caused by, say, atypical weather conditions. Due to the high stakes involved in many areas of ML decision-making, and the gravity of the damage that can be caused, this option seems unacceptable, as it implies practically abandoning the assignment of responsibility. It may be asked whether it is unfair to use autonomous systems if no one can take moral responsibility for harmful decisions. Some authors have taken the failure to assign moral responsibility as an argument in favor of completely abandoning the use of autonomous systems (Sparrow 2007; Asaro 2012).

Determining how to assign moral responsibility for harm done by ML decision-making thus seems extremely difficult. This is why a more practicable approach may be taken in the absence of an ethical resolution, or until the issue of moral responsibility is resolved. It might be practically more useful to try to define liability, even if moral responsibility is not sorted out. We will now present possible directions in resolving the issue of liability, which bypass the problems of finding morally responsible agents.

4.2 'Liability first'

As things currently stand, there are no established professionally or legally endorsed liability mechanisms for the regulation of the AI sector. The development of AI technology is proceeding faster than the development of legal and regulatory mechanisms to control and supervise it (Mittelstadt 2019). We thus propose potential solutions which might be introduced from regulatory practices of other sectors.

One possible direction would be to enable end-users to participate in the decision-making process by presenting

them with the potential risks and benefits of using ML and letting them consent to such a decision-making system being applied to their case. This solution would be similar to the informed consent practice used in other fields. For example, in medicine, informed consent means that patients are presented with the risks of a procedure and that they can consent to take those risks. By that means, informed consent is a type of involving the user in a decision about their case. The user thus takes on part of the responsibility for the risks the decision-making carries.

While this option could provide the basis for expedient regulation, it should be borne in mind that because of the inherent characteristics of ML, the informed consent practice may face certain challenges that do not arise in other fields. As already mentioned, the epistemic characteristics of ML often complicate the detection of errors and the attribution of responsibility for those errors. This involves determining whether the harm was a consequence of recklessness or abuse, rather than of an accidental error. In other fields, the cases of such unethical actions are not covered by informed consent, and the liability cannot be disclaimed. Since ML often precludes differentiating the two classes of cases, adopting the informed consent practice would mean considering all cases of harm to be unproblematic by default. In this way, the practice of informed consent bears the risk of being used as an alibi for human negligence or abuse.

Another direction in regulating liability in ML would be to proclaim that the liability is borne by the institution which benefits by either selling the ML software or by using it. The regulation would specify the types and amounts of compensation afforded to the affected party by the institution in cases of harm caused by ML decision-making. Such regulation already exists in other fields, in the form of the court ordering an institution to pay compensation or damages to the person for the harm suffered, known in many countries as the no-fault compensation system (Henry et al. 2015). This, again, does not require assigning moral responsibility for resolving liability. This solution has great practical advantages because it does not require finding an individual culprit or the exact source of the error that led to the detrimental outcome. This makes it especially convenient for the context of ML where it is incredibly difficult to determine this. No-fault liability would require that the end-user shows that the ML error was a causative factor in the resultant detrimental outcome, irrespective of who specifically is to blame. Thus, no-fault liability demands "proof of causation rather than proof of fault" (Gaine 2003). This direction of assigning liability to companies or institutions might seem more fair to the general public than the informed consent practice because liability is born by someone who profits from the use of ML, instead of someone who is negatively affected by it.

It should be noted that this solution comes with its own difficulties, albeit they do not arise from the inherent characteristics of ML, but rather from more general societal factors. Namely, AI technology promises to be tremendously profitable for those who own it. If the profit heavily outweighs the compensations, the penalties may not have noticeable practical significance, and the agents labeled liable might have no financial incentive to uphold any ethical principles when designing or using ML systems.

Legal solutions, however, desirable and urgent, without an ethical basis, do not seem to be the ideal solution and leave room for concern about the direction in which the AI sector will develop. The decisions about the directions of development and the kinds of AI technology are made mostly by private companies, not with the public interest in mind, but rather guided by market logic. In combination with the already mentioned increasing massivity of the use of ML in diverse areas of life, and the impact this decision-making has on human lives, the AI sector bears enormous power to shape the future of humanity. The unresolved issue of moral responsibility thus remains significant and should not be dismissed even if the legal regulation is well-established.

5 Conclusion

Using ML models for decision-making has important advantages over human decision-makers. The semi-autonomy of the model in processing large amounts of data, and the capacity to process the quantities which by far transcend human cognitive abilities help make ML models highly efficient and accurate in their predictions. On the downside, when faulty ML processes cause harm to the subjects of decisions, assigning responsibility for the harm faces serious difficulties, due to the unavoidable and inherent properties of ML systems that make them immune to human control and surveillance.

It seems that the further development of the ML field is directed toward ubiquitous automated decision-making on a massive scale in a variety of contexts, and that the problems with control and responsibility are more likely to be amplified in the future. Given the accelerated development of the field, resolving the regulation problems seems urgent to avoid possible abuse, and protect all actors involved in the decision-making.

To sum up, we uncover a conflict within the field of machine learning. What was considered its main advantage (handing over enormous quantities of information-processing tasks to a 'machine', which can perform them incredibly faster than humans) has turned out to be precisely the aspect that produces ethical problems for which it is not clear how they could be overcome while keeping the advantages. It is an open question whether the advantages of the use of ML are worth

the potentially irresolvable issues of moral responsibility in cases of harm caused by the decisions, and whether it is advisable to abandon the use of ML models for high-stakes decisions (Sparrow 2007), and go back either to human decision-making, or at least to simpler, transparent decision-making models (the latter was suggested by Rudin (2019) based on a different, but complementary rationale). Further research is needed to weigh the costs and benefits of such a 'regress'.

Acknowledgements We would like to thank Nenad Filipović for the engaged discussion and helpful comments on the early versions of this paper.

Data availability Non applicable.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ananny M, Crawford K (2018) Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc 20(3):973–989

Anderson M, Anderson S (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28(4):15–26

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. Retrieved November 9, 2021, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Apps P (2021) New era of robot war may be underway unnoticed. Reuters. Retrieved September 7, 2021, from https://www.reuters.com/article/apps-drones-idUSL5N2NS2E8

Asaro PM (2012) On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. Int Rev Red Cross 94(886):687–709

Asaro PM (2014) A body to kick, but still no soul to damn: legal perspectives on robotics. In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, pp 169–186

Boge FJ, Grünke P (2019) Computer simulations, machine learning and the Laplacean demon: opacity in the case of high energy physics. In: Kaminski A, Resch M, Gehring P (eds) The science and art of simulation II. Springer

Bryson JJ (2010) Robots should be slaves. In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issues. John Benjamins, pp 63–74

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data &Society, 3(1). https://doi.org/10.1177/2053951715622512

Butler D (2016) Tomorrow's world. Nature 530:399–401

Cornock M (2011) Legal definitions of responsibility, accountability and liability. Nurs Child Young People 23(3):25–26

Fischer JM, Ravizza MSJ (1998) Responsibility and control: a theory of moral responsibility. Cambridge University Press

Flores AW, Lowenkamp CT, Bechtel K (2016) False positives, false negatives, and false analyses: a rejoinder to "Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks." Fed Probat J 80(2):38–46

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28:689–707

Gaine WJ (2003) No-fault compensation systems. BMJ 326(7397):997–998

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of the 2018 IEEE 5th international conference on Data Science and Advanced Analytics (DSAA). IEEE, pp 80–89

Goertzel B (2002) Thoughts on AI morality. Dyn Psychol Int Interdiscip J Complex Ment Process. Retrieved October 31, 2021, from http://www.goertzel.org/dynapsyc/2002/AIMorality.htm

Goh YC, Cai XQ, Theseira W, Ko G, Khor KA (2020) Evaluating human versus machine learning performance in classifying research abstracts. Scientometrics 125:1197–1212

Goodman B, Flaxman S (2017) EU regulations on algorithmic decision-making and a 'Right to Explanation.' AI Mag 38(3):50–57

Grossmann J, Wiesbrock HW, Motta M (2021) Testing ML-based systems. Federal Ministry for Economic Affairs and Energy. https://docbox.etsi.org/mts/mts/05-CONTRIBUTIONS/2022/MTS(22)086017_Testing_ML-based_Systems.pdf

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput Surv 51(5):1–42

Gunkel DJ (2020) Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf Technol 22:307–320

Hall JS (2001) Ethics for machines. Kurzweil Essays. Retrieved June 15, 2021, from KurzweilAI.net http://www.kurzweilai.net/ethics-for-machines

Hanson FA (2009) Beyond the skin bag: on the moral responsibility of extended agencies. Ethics Inf Technol 11:91–99

Hart E (2019) Machine learning 101: the what, why, and how of weighting. KDnuggets. Retrieved May 21, 2021, from https://www.kdnuggets.com/2019/11/machine-learning-what-why-how-weighting.html

Henry LM, Larkin ME, Pike ER (2015) Just compensation: a no-fault proposal for research-related injuries. J Law Biosci 2(3):645–668

Hoffman RR, Mueller ST, Klein G, Litman J (2018) Metrics for explainable AI: challenges and prospects. XAI Metrics. Retrieved October 1, 2021, from https://arxiv.org/ftp/arxiv/papers/1812/1812.04608.pdf

Humphreys P (2004) Extending ourselves: computational science, empiricism, and scientific method. Oxford University Press

Humphreys P (2009) The philosophical novelty of computer simulation methods. Synthese 169:615–626

Johnson DG (2006) Computer systems: moral entities but not moral agents. Ethics Inf Technol 8(4):195–204

Johnson DG, Miller KW (2008) Un-making artificial moral agents. Ethics Inf Technol 10(2–3):123–133

Lauret J (2019) Amazon's sexist AI recruiting tool: how did it go so wrong? Medium. Retrieved November 9, 2021, from https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e

Lee J (2020) Is artificial intelligence better than human clinicians in predicting patient outcomes? J Med Internet Res 22(8):e19918. https://doi.org/10.2196/19918

Lipton ZC (2016) The mythos of model interpretability. In: 2016 ICML workshop on human interpretability in machine learning (WHI 2016). New York. https://arxiv.org/abs/1606.03490

Lunney M, Oliphant K (2013) Tort law, 5th edn. Oxford University Press

Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6(3):175–183

McKenna M (2008) Putting the lie on the control condition for moral responsibility. Philos Stud 139:29–37

Mehta S (2022) Deterministic vs stochastic machine learning [Blog post]. https://analyticsindiamag.com/deterministic-vs-stochastic-machine-learning/
Miller T (2017) Explanation in artificial intelligence: insights from the social science. Artif Intell 267:1–38
Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507
Mittelstadt B, Russell C, Wachter S (2019) Explaining explanations in AI. In: FAT* '19: conference on fairness, accountability, and transparency (FAT* '19). Retrieved October 30, 2021, from https://arxiv.org/pdf/1811.01439.pdf
Molnar C (2019) Interpretable Machine Learning. Available online: https://christophm.github.io/interpretable-mlbook/
Moor J (2006) The nature, importance and difficulty of machine ethics. IEEE Intelligent Systems 21(4): 18-21
Mowshowitz A (2008) Technology as excuse for questionable ethics. AI Soc 22(3):271–282
Nissenbaum H (1996) Accountability in a computerized society. Sci Eng Ethics 2(1):25–42
Ombach J (2014) A short introduction to stochastic optimization. Schedae Informaticae 23:9–20
Paez A (2019) The pragmatic turn in explainable artificial intelligence (XAI). Mind Mach 29:441–459
Pant K (2021) AI in the courts [Blog post]. Retrieved from https://indianexpress.com/article/opinion/artificial-intelligence-in-the-courts-7399436/
Price M (2019) Hospital 'risk scores' prioritize white patients. Science. Retrieved November 9, 2021, from https://www.science.org/content/article/hospital-risk-scores-prioritize-white-patients
Ribera TM, Lapedriza A (2019) Can we do better explanations? A proposal of user-centered explainable AI. In joint Proceedings of the ACM IUI 2019 workshops
Rudin C (2019) stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
Russ M (2021) Artificial intelligence, machine learning, and deep learning—what is the difference and why it matters [Blog post]. Retrieved from https://bluehealthintelligence.com/how-to-know-the-difference-between-artificial-intelligence-machine-learning-and-deep-learning-and-why-it-matters/
Russell SJ, Norvig P (eds) (2016) Artificial intelligence: a modern approach. Pearson Education Limited, Cham
Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR (eds) (2019) Explainable AI: interpreting, explaining and visualizing deep learning. Springer
Schembera B (2017) Myths of Simulation. In: Resch MM, Kaminski A, Gehring P (eds) The science and art of simulation I: exploring—understanding—knowing. Springer, Cham, pp 51–63
Sidelov P (2021) Machine learning in banking: top use cases [Blog post]. Retrieved from https://sdk.finance/top-machine-learning-use-cases-in-banking/
Siponen M (2004) A pragmatic evaluation of the theory of information ethics. Ethics Inf Technol 6(4):279–290
Sparrow R (2007) Killer robots. J Appl Philos 24(1):62
Srećković S, Berber A, Filipović N (2022) The automated Laplacean demon: how ML challenges our views on prediction and explanation. Mind Mach. https://doi.org/10.1007/s11023-021-09575-6
Sullins JP (2006) When is a robot a moral agent? Int Rev Inf Ethics 6(12):23–30
Talbert M (2022) Moral responsibility. In: Zalta EN, Nodelman U (eds) The Stanford encyclopedia of philosophy (Fall 2022 edition). https://plato.stanford.edu/archives/fall2022/entries/moral-responsibility/
Tkachenko N (2021) Machine learning in healthcare: 12 real-world use cases to know [Blog post]. Retrieved from https://nix-united.com/blog/machine-learning-in-healthcare-12-real-world-use-cases-to-know/#:~:text=One%20of%20the%20uses%20of,decision%2Dmaking%20and%20patient%20care.
Turing A (1999) Computing machinery and intelligence. In: Meyer PA (ed) Computer media and communication: a reader. Oxford University Press, pp 37–58
UNI Global Union (2018) 10 principles for ethical AI. UNI Global Union, February 21, 2021. http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/
Varshney KR, Alemzadeh H (2017) On the safety of machine learning: cyber-physical systems, decision sciences, and data products. Big Data 5(3):246–255
Verbeek PP (2011) Moralizing technology: understanding and designing the morality of things. University of Chicago Press
Wachter S, Mittelstadt B, Floridi L (2016) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Privacy Law 7(2):76–99
Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv J Law Technol 31(2):841–887
Wallach W, Allen C (2009) Moral machines: teaching robots right from wrong. Oxford University Press
Wang F, Rudin C, McCormick TH, Gore JL (2019) Modeling recovery curves with application to prostatectomy. Biostatistics 20(4):549–564
Wang H, Shuai P, Deng Y et al (2022) A correlation-based feature analysis of physical examination indicators can help predict the overall underlying health status using machine learning. Sci Rep 12:19626
Wexler R (2017) When a computer program keeps you in jail: how computers are harming criminal justice. New York Times. Retrieved October 3, 2021, https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html
Wyber R, Vaillancourt S, Perry W, Mannava P, Folaranmi T, Celi LA (2015) Big data in global health: improving health in low- and middle-income countries. Bull World Health Organ 93(3):203–208
Yampolskiy R (2020) Unexplainability and incomprehensibility of AI. J Artif Intell Conscious 7(2):277–291
Yeung K (2019) Responsibility and AI: a study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Council of Europe Study Series. Council of Europe
Zednik C (2019) Solving the black box problem: a normative framework for explainable artificial intelligence. Philos Technol 34:265–288
Zerilli J, Knott A, Maclaurin J, Gavaghan C (2019) Transparency in algorithmic and human decision-making: is there a double standard? Philos Technol 32:661–683
Zhao T, Dai E, Shu K, Wang S (2022) Towards fair classifiers without sensitive attributes: exploring biases in related features. In: Conference: WSDM '22: the fifteenth ACM international conference on web search and data mining, pp 1433–1442. https://doi.org/10.1145/3488560.3498493
Zimmerman MJ (1997) Moral responsibility and ignorance. Ethics 107(3):410–426