THEORIA 3
(2024) : 67 : str. 177-192

https://doi.org/10.2298/THEO2403177B
Originalni naučni rad
Original Scientific Paper

Andrea Berber & Jelena Mijić

UNDERSTANDING MORAL RESPONSIBILITY IN AUTOMATED DECISION-MAKING: RESPONSIBILITY GAPS AND STRATEGIES TO ADDRESS THEM

SUMMARY: *This paper delves into the use of machine learning-based systems in decision-making processes and its implications for moral responsibility as traditionally defined. It focuses on the emergence of responsibility gaps and examines proposed strategies to address them. The paper aims to provide an introductory and comprehensive overview of the ongoing debate surrounding moral responsibility in automated decision-making. By thoroughly examining these issues, we seek to contribute to a deeper understanding of the implications of AI integration in society.*

KEYWORDS: *Moral responsibility, Responsibility gap, Problem of many hands, Automated decision-making, Explainable AI, Machine learning, Artificial intelligence*

Introduction

The questions surrounding the integration of Artificial Intelligence (AI) into human society are garnering significant attention from both the general public and academic communities. This attention is driven not only by the awe-inspiring capabilities of certain AI systems but also by concerns about their potentially harmful impacts and the significant ethical questions they raise. This paper aims to delve into the utilization of machine learning-based systems in decision-making processes and examine ethical concerns surrounding them. Particularly, it addresses the impact of using ML decision-making systems on human responsibility, traditionally understood. A primary focus will be on the emerging responsibility gaps and potential strategies to address them. The paper unfolds in the following manner: In Section 1, we present introductory insights into the utilization of ML in decision-making. Section 2 delves into the concept of moral responsibility, offering a comprehensive conceptual framework. In Section 3, we elucidate the emergence of responsibility gaps and explore their various types. Lastly, Section 4 provides commentary on different proposals aimed at addressing these responsibility gaps.

1. ML in high-stakes decision-making

The integration of ML in decision-making processes has become increasingly prevalent across various critical fields, revolutionizing the way choices are made in legal systems, social benefits, human resources, banking, warfare, and healthcare. Decisions influenced by ML models in these domains have the potential to significantly impact individuals' lives, ranging from unjustly denying opportunities to endangering health or causing human rights violations. This phenomenon is commonly referred to as *high-stakes decision-making*, and it will be the main focus of this paper.

To illustrate, we will provide a few examples of the ML decision-making influence. In legal systems, ML algorithms are employed to make jail sentence recommendations and assess parole eligibility, raising concerns about fairness and justice. The widely used Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) model in the U.S. justice system has faced scrutiny for reported racial biases, highlighting the ethical implications of ML-driven decisions in the legal domain. Social benefit programs such as housing and social stamps eligibility also leverage ML for decision-making. The potential consequences of inaccurate determinations include denying critical support to those in need, reflecting the profound impact of ML in this field. Human resources departments utilize ML to streamline job opportunities. Yet issues of bias have arisen, as exemplified by Amazon's recruiting tool being accused of gender bias. In banking, ML systems assess loan opportunities, introducing efficiency but also potential risks of biases. Meanwhile, in healthcare, diagnostic and therapy recommendations powered by ML, like IBM's Watson for Oncology, have faced criticism for unsafe recommendations, raising concerns about the impact on patient well-being. Arguably, the most prominent illustration of potential harm resulting from the influence of ML-based technologies is the deployment of AI-based weapons in warfare, which can lead to violations of *jus in bello* and consequently to widespread devastation and the loss of human lives.

Before delving into the question of moral responsibility for decisions and outcomes influenced by ML, we need to make a few preliminary clarifications. One key consideration is the extent of ML's influence in the decision-making process. Firstly, we can use the ML model to reach a final decision without human involvement, such as selecting three promising job candidates from a pile of received CVs. Although a human HR manager would implement this decision by calling these candidates for a job interview, the selection process would be *entirely ML-driven*.

The second option involves *combining human expertise with ML*. In this scenario, the ML model recommends ten candidates, and then a human, drawing on their expertise, chooses the three most promising ones. It's important to note that in this collaborative decision-making option, the levels of human involvement may significantly vary. A human might use the ML model to exclude unqualified candidates and then

proceed with decision-making independently. Alternatively, a human could review ML decisions for potential mistakes and biases in the model, adding a layer of oversight to ensure fairness and accuracy. In this paper, we are primarily interested in the first option because the second one involves many potential levels of human involvement, which, in turn, can affect the way responsibility is assigned. Considering all these nuances surpasses the scope of this research. Nonetheless, we will explore collaborative decision-making or so-called *human-in-the-loop* as a strategy to address the challenges posed by the first option. Consequently, we will delve into certain complexities associated with this approach.

Yet another point we want to stress is the type of opacity or black-boxiness we are interested in. Some ML models are opaque due to proprietary rights, even though they may not be inherently opaque. Additionally, we can discuss models being opaque to the recipients of decisions or users due to a lack of expertise (Burrell 2016). However, we are specifically interested in the opacity that affects both designers and users, placing the data engineer and, for instance, a physician using the model in diagnostics in the same epistemic position (Smith 2021). This opacity arises from the fact that calculations performed by the models surpass human cognitive abilities in terms of complexity. While the first two types of opacity are, in principle, removable by subtracting proprietary rights and educating users of the systems, the last one is not, as it is related to the inherent limitations of human cognitive abilities.

2. Moral Responsibility - Conceptual Framework

The responsibility for the outcomes of autonomous AI systems' actions is a crucial aspect of discussions about trustworthy AI within the scientific and public communities. *Moral responsibility* (*accountability*) cannot be overlooked if we aim to develop and use AI in a responsible and accountable manner. In the previous sentences, we used the terms "responsibility" and "accountability" with different meanings and thus suggested varieties of responsibility. Philosophers often discuss the concept of moral responsibility in the sense of an agent's responsibility for their actions and the outcomes of those actions. This type of responsibility usually applies after the event has occurred, and is referred to as *backward-looking responsibility*. However, a complete analysis of backward-looking responsibility must include an analysis of *forward-looking* or *prospective responsibility* that applies when the agent is responsible for something that has not yet occurred.

Two types of forward-looking responsibility are *responsibility-as-virtue* and *responsibility-as-obligation* (van de Poel and Sand 2021: 4773). When someone is described as responsible in a virtue sense, it means that their character is positively evaluated, which is the opposite of being irresponsible)(Vincent 2009). Responsibility as

a virtue is not just about a single act. It involves having historical insight into an agent's personal quality or virtue and relates to actively assuming responsibilities. Among other things, it can mean the tendency of the agent to fulfill their duties and take responsibility for the consequences of their actions (Williams 2008). On the other hand, responsibility as a (moral) obligation refers not only to the tasks or obligations that come with one's social, institutional, and other roles but also to the *obligation to see to it that something is the case*.[1] For instance, a sea captain is not just navigating the ship or is in charge of her but is also responsible for ensuring the safety of the passengers and crew; he has to see to it that the passengers are transported safely. Following Goodin, van de Poel suggests that obligations should not be understood as individual actions but rather as *responsibilities*: general classes of activities that can be delegated. However, the captain certainly has a responsibility that requires some action on his part of a supervisory nature: he has to see to it that the passengers are safe. Although individuals have certain obligations they agree to undertake beforehand, they can also be held responsible retroactively (ie. in a backward-looking sense of responsibility) if they fail to discharge them. However, the opposite is not necessarily true. In other words, things for which an individual is *morally* responsible do not necessarily have to be things they are obliged to do (Oshana 1997: 72).

There is often confusion about the very concept of moral responsibility. This confusion arises not only due to the lack of differentiation between various kinds of responsibility but also due to the lack of differentiation between different analyses of moral responsibility. Traditionally, moral responsibility has been viewed as a unified concept. However, with the emergence of diverse theoretical positions[2] on the relationship between determinism and free will, which form the basis of considerations about moral responsibility a pluralistic view of moral responsibility has developed (Jeppsson 2022). The advocates of pluralism propose the idea of distinguishing between more and less demanding kinds of moral responsibility to ensure the continuation of attributing moral responsibility, even in situations where the traditional concept of moral responsibility based on desert is not appropriate (in the case of agents with deprived childhoods or agents with certain mental impairments, for example). However, this approach has confused the relationship between moral responsibility and blameworthiness, answerability, accountability, attributability, desert, and punis-

1 When considering varieties of responsibility, philosophers generally refer to the descriptive concept of role responsibility, which refers to the agent's tasks, or the descriptive concept of responsibility, which refers to having authority and being in charge of something. Noting that not every role defines a moral obligation, van de Poel includes Goodin's normative concept of *task responsibility* as an obligation to see to it that some state of affairs occurs. See: van de Poel 2015; van de Poel & Sand 2021; Goodin 1995.

2 We primarily refer to moral responsibility skepticism, a controversial stance that questions whether anyone deserves praise or blame, reward or punishment.

hment.[3] It is important to note that these approaches to moral responsibility are fairly recent innovations. We won't delve into all the nuances of these concepts, but it's important to acknowledge them to avoid misunderstandings.

When discussing moral responsibility in the context of AI systems, authors make use of several related terms, with "accountability" being the most common.[4] To avoid confusion, we rely on a philosophical and legal analysis of backward-looking responsibility. Although legal and moral concepts of responsibility are not identical, it's a common practice to draw a parallel between these two related concepts when considering responsibility for the consequences of actions.[5]

Investigation in this paper will rely on an understanding of moral responsibility that aligns with our ordinary assumptions of moral responsibility. These assumptions form the basis of current moral and legal practices. Specifically, we will explore moral responsibility as accountability and blameworthiness (This paper focuses on responsibility for bad actions, ie. culpability). Accordingly, we introduce the stance that the agent is morally responsible if they acted in a way that they deserve to be blamed or praised (or be punished or rewarded). Furthermore, the action is of a nature that it is appropriate to expect the agent to explain and provide an account and justification for why they did what they did when questioned about it (Oshana 1997: 76-77). In other words, the agent is accountable for their actions.

A moral agent is an agent who meets the requirements of a responsible agency. Philosophers often analyze the concept of responsible agency using the Aristotelian conditions. Aristotle, in the *Nicomachean Ethics*, described these conditions negatively as *ignorance* and *force*: "Since virtue is concerned with passions and actions, and on voluntary ones praise and blame are bestowed, on those that are involuntary pardon, and sometimes also pity [...] Those things, then, are thought involuntary, which take place by force or by reason of ignorance [...]" (*NE* 1110a). However, modern-day philosophers have formulated them positively as the *epistemic condition* and *control condition* (Fischer and Ravizza 1998: 13). Furthermore, we recognize these conditions when considering criminal responsibility: *actus reus* and *mens rea*.

The control condition is commonly referred to as the *freedom condition*. According to this view, free will is "whatever ability is required of a morally responsible agent to have sufficient control over her culpable conduct that she would deserve blame for it" (McKenna 2022: 28). Thus, control condition or *actus reus* requires that there must be a causal connection between the agent and the outcome. This means that the agent's action or omission must have led to the outcome. The action must be vo-

3 See: Watson 1996, Shoemaker 2015.

4 For a detailed analysis of the different ways in which the term "accountability" is used in this context, see: Lechterman 2022.

5 See: Feinberg 1962.

luntary, which means that the person must have control over it (that the action belongs to the person in a relevant sense). However, it is important to distinguish between the control condition and the concept of *causal responsibility*. An agent can be held causally responsible for an outcome but be acting under the influence of coercion or manipulation. Causal responsibility does not necessarily imply that the agent is morally or legally responsible for their actions (Feinberg 1970: 188). This is because there may be situations where the agent's actions cannot be considered their own (their body movements may not be counted as instances of their actions), which may prevent them from being blamed or praised. It is not uncommon for a person to be causally linked to an outcome yet not be morally or legally responsible for it. When an agent's action is caused by a psychoactive substance or a mental illness, such as paranoid schizophrenia, it may not be a voluntary action. In such cases, the action does not arise from the agent's intentional choice.

Therefore, the epistemic condition or *mens rea*[6] requires that a particular state of mind accompany an action. It means that the agent must be self-aware, conscious of the act and the circumstances in which they act, and able to recognize moral or legal norms[7]. Furthermore, the agent must be sensitive or responsive to reasons for actions (Fischer and Ravizza 1998).

Accordingly, law and philosophy do not recognize, for example, children up to a certain age and people with some mental disorders as moral agents. Because they cannot satisfy the conditions for responsible action, neither animals nor machines are considered moral agents. Therefore, moral responsibility cannot be attributed to any of them. Given the current technological state of affairs, AI also does not meet the conditions for moral agency, which motivates special issues related to moral responsibility in the context of automated decision-making.

3. Responsibility gaps and how they arise

Contemporary society, characterized by technological development and innovation, is facing a new problem: the so-called *responsibility gap*. (de Jong 2020, Danaher 2016, Sparrow 2007, Matthias 2004). At times, it may seem necessary to assign responsibility for a particular outcome, especially in high-stakes contexts, as explained in Section 1. The responsibility we are seeking is not merely causal but also moral and legal as explained in Section 2. However, it is possible that the appropriate party to hold responsible

6 The law recognizes four types of mens rea: specific intent, knowledge, recklessness, and negligence.

7 Legal and moral responsibility are not identical concepts and are not always applied in the same situations. A person may be legally responsible for an action because it is against the law but may not be morally responsible if the action is not considered immoral.

may not be immediately apparent. In other words, it may not be clear or justifiable to assign responsibility to any particular individual for a given outcome.

The reason for the responsibility gap concerning certain AI-based technologies lies in their design to be (technically) autonomous, meaning they navigate through informational or physical space without direct human control (Matthias 2004). Thus, whether the individual is a designer or a person using the machine, they lack control over machine actions and outcomes. In the case of algorithmic decision-making systems based on ML, the system itself extracts or learns decision-making rules from the data it is trained on (Srećković *et al*. 2023, Berber and Srećković 2023). This implies that responsibility for the outcome or decision, cannot be attributed to a human agent because they do not have relevant control over it. The risks of failure to meet the control condition are especially evident in the case of artificial systems designed to operate fully autonomously (without human intervention), such as lethal autonomous weapons (LAWS).

However, when it comes to combining human expertise and AI-based systems in a so-called *human-in-the-loop* situation, it could seem that if the human agent has the final say in the decision and meets the control condition, they are therefore morally responsible. Nonetheless, in this situation, we encounter the failure to satisfy another condition for moral responsibility - the epistemic condition. A human working with an AI system lacks insight into the internal workings of the system, making it difficult to assess the justification or validity of the recommendation the system provides (Baum *et al*. 2022, Berber and Srećković 2023). We will further discuss how this kind of opacity affects the assignment of responsibility in Section 4. For now, we can say that the responsibility gap arises when neither the human agent (designer, programmer, operator) nor the AI-based system meets the conditions for moral responsibility. As explained in Section 2, although an AI-based system possesses a certain causal contribution to the decision-making process (indeed, this causal impact impedes relevant human control), it cannot satisfy the epistemic condition due to a lack of attributes such as consciousness and self-awareness.

The problem of ascribing responsibility for AI-based decision-making and its consequences is further amplified by the fact that many different actors are usually involved in reaching a decision. This problem is not unfamiliar in philosophy and is commonly referred to as *the problem of many hands*. In a nutshell, the problem of many hands is the problem of attributing, or properly distributing, individual responsibility in collective environments:

> "Because many different officials contribute in many ways to decisions and policies of government, it is difficult even in principle to identify who is morally responsible for political outcomes. This is what I call the problem of many hands" (Thompson 1980: 905).

Thus, this problem refers to situations where multiple agents have contributed to a procedure that resulted in a negative consequence, and it's not clear who should be held morally responsible. It is not just an epistemological problem that stems from the situation's complexity and the difficulty of determining who possessed relevant control and knowledge (van de Poel 2015; Lechterman 2022). In some cases, the problem could not be reduced by requiring better record-keeping because it could be the case that no one individually met the conditions for responsibility for a specific consequence. Alternatively, the situation could be such that everyone fulfilled their obligations, but a negative consequence still arose, as a matter of complex interaction of individual effects. It may also be established that some individuals did not fulfill their obligations, but the damage caused is disproportionate to their individual responsibility (List 2022: 133). While the problem of many hands originates in political philosophy, it takes on a new dimension in the context of information technology. Not only are AI decisions ultimately the product of numerous agents (developers, designers, operators, users), but also the non-human agents.

The subject of responsibility gap(s) generated by AI is a relatively recent development, and consequently, various meanings of the term "responsibility gap" have emerged. For example, in the European Commission's report on the ethics of autonomous vehicles, five techno-responsibility gaps were identified: "a (*public) accountability* gap (no one to give public account for the harm done), a *culpability gap* (no one to morally blame for the harm), a *compensation gap* (no one to pay damages for the harm), an *obligation gap* (a danger that no one has the duty to ensure that the machine avoids or minimizes harm), and a *virtue gap* (no one tries to cultivate a disposition to take responsibility for the actions of machines) (Danaher 2022: 4). This is not the only available classification framework for different types of gaps, but it aligns with our initial analysis of the types of responsibilities.[8] Varieties of responsibility gaps can be understood by considering the different kinds of responsibilities that are emphasized depending on the purpose for which the responsibility is assigned, for example, due care for others, remediation, maintaining moral community, or retribution (van de Poel 2015: 20). As will be shown, these frameworks reveal the complexity of the responsibility gap problem and highlight flaws in some of the attempts to address it.

This complexity can be illustrated by pointing to the insufficiently comprehensive solutions to the responsibility gap. For example, the compensation gap may arise due to the inconsistency of the existing legal standards for determining the compensation giver and the emergence of new technology such as autonomous machines that may cause damage. This gap can be bridged by alternative legal standards. Those might be *strict liability* standards that do not include fault and can be passed from one party to another (Danaher 2016). By appealing to retributivist intuitions, however, Danaher

8 For a similar framework see: Santoni de Sio and Meccaci 2021.

suggests this still seems to leave the culpability gap wide open. Danaher's argument refers to the psychological evidence that people are innate retributivists, so "when they are harmed or injured they look for a culpable wrongdoer who is deserving of punishment" (Danaher 2016: 299). Since restitution and compensatory damages are not intended to be punitive, as their main objective is to reinstate the victim to their pre-injury condition rather than to cause harm to the offender, this approach does not address the culpability gap. The general takeaway point in the complex landscape of gaps is that some gaps may be easier to bridge than others, and closing one type of gap may still leave others open.

4. How to deal with responsibility gaps

4.1 Elimination of black-boxes

The authors who initially explored the problem of the responsibility gap recommended reducing or eliminating the usage of autonomous systems due to their unpredictable nature.[9] The rationale behind this approach is that responsibility gaps cannot be bridged, and in some contexts, we simply cannot afford to lose the *locus* of responsibility. Therefore, we should consider giving up the use of autonomous systems. However, the radical proposal to eliminate the use of autonomous systems faces several types of objections. Firstly, it seems unjustified to insist on eliminating their use in low-stake or low-risk contexts based on the perceived risk of their use in high-stake contexts (Lechterman 2022). Second, given that autonomous systems are already in use, the question is whether it is realistic to expect their prohibition. Take, for example, the context of warfare. According to the March 2020 UN report on the military conflict in Libya, the *Kargu-2* attack drones (LAWS based on ML and image processing) that can be operated manually or autonomously "were programmed to attack targets without requiring data connectivity between the operator and the ammunition: in effect, a true 'fire, forget and find' capability" (United Nations 2021). Therefore, it can be argued that LAWS are already being used. Hence, any attempt to prohibit them, as Lechterman suggested, must consider the potential for prisoners' dilemmas that can hinder the intended outcomes of such a ban. The possibility of black markets should also be considered (Lechterman 2022).

The decision to use ML in some of the critical fields is driven by its speed, efficiency, and capacity to process vast amounts of data. ML systems excel in situations where timely decisions are crucial, such as in healthcare, financial fraud detection, and justice systems. The ability to process large datasets and consider numerous parameters, including the latest evidence from clinical studies, contributes to the efficiency of ML systems

9 See: Matthias 2004; Sparrow 2007.

in decision-making. ML's capability to surpass human abilities is a compelling factor. In some areas, ML systems demonstrate higher accuracy and fewer mistakes than human counterparts, such as in screening 3D radiological images for cancer cells (McKinney *et al*. 2020). Also, a study shows that AI systems match or surpass physician diagnostics in a range of medical specialties and even in some parameters for conversation quality, such as politeness (Tu *et al*. 2024). Economic reasons, including profits for software development companies and cost-effectiveness in specific application fields, further drive the adoption of ML in decision-making processes.

It has also been proposed that we should stop using automated systems based on black-boxed or non-transparent models for high-stakes but switch to simpler, interpretable models, at least for high-stakes decision-making (Rudin 2019). Since the opacity of the model's functioning is deemed as a property that significantly aggravates (and may even absolutely prevent) discerning, correcting, and explaining mistakes (Berber and Srećković 2023), the solution may be sought in using transparent models. This would imply not abandoning automated decision-making altogether, but rather those based on black box algorithms. However, when it comes to the option of banning a new technology, it is always necessary to take into account how implementable this option is in practice and what the balance of losses and gains from this option would be.

Doubtful of solving the responsibility gap problem by eliminating autonomous machines, authors focus on finding ways to avoid unacceptable risks and assign responsibility for AI-caused harms. The following two proposed solutions are along these lines.

4.2 Keep human in the loop

Some researchers advocate that we cannot afford to lose the *locus* of responsibility when it comes to high-stakes decisions. According to this approach, we should always maintain human oversight in high-stakes contexts. This means that there will always be a designated individual who holds direct responsibility for making ultimate decisions informed by recommendations from automated systems (Baum *et al*. 2022). However, this approach requires that automated decision-making systems offer reason-based explanations similar to the ones humans use to justify their decisions, thereby making them understandable to humans working with the system. Having this kind of explanation would facilitate informed judgments by humans regarding whether to endorse or dismiss the system's recommendations. The important feature of this solution is that it aims to provide the necessary degree of explainability to satisfy the epistemic condition for moral responsibility.

However, this solution has some important challenges that we want to highlight. First, it depends on the success of the Explainable AI (XAI) project, which, in a nutshell, is an attempt to unpack black box models and provide sufficient understanding of the-

ir inner workings to end users. This approach demands specific kinds of explanations that are reason-based and thereby susceptible to human assessment. It is a question of technical feasibility whether it is possible to provide explanations that satisfy this demand and are simultaneously faithful to the original black box model calculations. Second, putting human beings in charge of overseeing the automated system may diminish some of the initial advantages and could nullify the advantages of employing autonomous systems in the first place. For instance, in medical diagnostics, the imperative for speed and efficiency could be critical for saving lives. Introducing a human supervisor to oversee an automated system could compromise efficiency and undermine the initial justification for implementing the system (Berber 2023).

Furthermore, concerns arise regarding the moral and epistemic position of the human involved. How would the presence of an "automated advisor" impact the moral and epistemic responsibility of humans? In instances of disagreement with the system, to what extent is it rational for humans to adhere to their judgment, particularly if the system demonstrates a considerably lower error rate than humans (Smith *et al*. 2023, Berber 2023)? It should be reasonably expected that humans are able, in real-time and complex decision-making situations, to timely identify and correct system errors. However, as argued by Elish (2019) based on the case studies of the Three Mile Island accident in 1979 and the crash of Air France in 2009, engineers and managers of different systems often create "contrary dynamics", which implies that automation is generally safer and superior, but humans are superior in cases where something goes against plans. Expecting a human to successfully jump into an emergency at the last minute may be unrealistic and unfair. This creates the situation Elish calls the "moral crumple zone," misattributing responsibility to human actors who had limited control over automated autonomous systems.

Overall, this approach hinges on the need to design automated systems that will be helpful to human agents (surpassing some of the human weaknesses) but remain understandable, correctable, and controllable.

4.3 Design systems that are under meaningful human control

As we saw in Section 3, losing control over automated systems is the factor that prevents ascribing responsibility to human agents, thereby opening responsibility gaps. Thus, one way to solve this situation would be by designing systems that would remain under human control. The idea is to overcome the shortcomings of seemingly one-sided proposals to bridge the responsibility gap, for example, specific legal regulations or strictly technical solutions (Santoni de Sio and Mecacci 2021: 1072). Although said proposals are not insignificant and certainly play a valuable role in the discussion on the development and use of AI, as Santoni de Sio and Meccaci note, they overlook active responsibility gaps (obligation gap and virtue gap). In the case

of innovations and new technologies, it takes time to determine the balance between conflicting values and moral standards (van de Poel and Sands 2021: 4776), which implies that agents involved in the development and use of AI may not be able to identify and recognize relevant values, interests, and reason that should shape their obligations. The obligation gap enables manipulation of the attribution of moral responsibility, which further motivates unwillingness to assume and take responsibility and creates a virtue gap. A more comprehensive proposal would have to anticipate ways of bridging the multiple aspects of the responsibility gap, especially those associated with forward-looking responsibility and whose connection to the culture of responsibility is the most immediate.

This rationale underpins the concept of *meaningful human control* that has emerged in the legal and political debate surrounding the ethical concerns of employing LAWS. Santoni de Sio and van den Hoven (2018) elaborate on the concept of meaningful human control and propose that autonomous systems should be designed to implement it. Drawing from a philosophical analysis of moral responsibility they identify two key conditions for meaningful human control. The first condition is the *tracking condition* which requires that the system track or be responsive to reasons (including values and intentions) of relevant human agents. On the theoretical side, this necessitates identifying the relevant human agents and reasons in different scenarios where the system is applied and, on the technical side, designing the system with the required level of responsiveness to those reasons under various circumstances. The second, *tracing condition* stipulates that the system should be designed in a manner that allows for tracing the outcome of its operations back to at least one human involved in its design and operation. To meet the tracing condition it is necessary to ensure that various humans throughout the chain possess the technical skills and psychological readiness to fulfill their obligations as well as to comprehend their role in and responsibility for the behavior of the autonomous system. Furthermore, this requires the establishment of new normative frameworks, such as updated legal regulations for assigning liability in cases of harm involving autonomous systems. Alongside, as emphasized by the authors, it is imperative to develop adequate educational and training programs aimed at enhancing understanding of these systems' operations, as well as the associated risks and responsibilities involved in their design and operation.

The idea behind this proposal is to preserve morally relevant control within the context of AI systems by ensuring that the system aligns with the *reasons* and *capacities* of human agents. This proposal aligns with traditional assumptions about moral responsibility, keeping humans as the sole moral agents and responsible parties. This aspect could be seen as an advantage, as it may lead to easier acceptance and implementation in society. Additionally, this proposal is comprehensive as it considers various aspects of the overall complex socio-technological situation, such as moral, technical, legal, psychological, and educational factors. These elements need to be

aligned to ensure that we design and implement systems under meaningful human control.

The broad scope of this proposal may also pose an obstacle to its implementation. Firstly, certain theoretical controversies could impede progress, such as the precise determination of the human agent responsible, which could lead us back to the many hands problem. Additionally, there might not be universal agreement on the moral values to be integrated into the system and even if consensus is reached on a particular value, disagreements over its exact interpretation may arise. Also, implementing this solution necessitates multidisciplinary dialogue among experts from diverse backgrounds and their active involvement in system design, a practice not currently commonplace. Additionally, such a solution could slow down technological development because meaningful human control, along with all the aforementioned challenges, must be ensured during the actual design process rather than after the fact, and we may wonder if this is a realistic expectation.

Concluding Remarks

In this paper, we aimed to offer an overview of the debate surrounding the responsibility gaps concerning autonomous machines. We focused on illustrating the complexity of the responsibility gap issue and highlighted various types of responsibility gaps, reflecting the intricate socio-technological context of this matter. Regarding potential solutions, we discussed the most prominent ones, without claiming to provide an exhaustive catalog, and refrained from asserting which one is superior. Our approach was to examine each solution from both positive and negative angles. In this paper, we addressed responsibility gaps as a problem requiring attention. However, contrary to authors who emphasize finding ways to bridge and close the responsibility gap (in one form or another), some suggest that "techno-responsibility gaps are, sometimes, to be welcomed," and one of the advantages of autonomous machines is that they allow us to embrace certain types of responsibility gaps (Danaher 2022). Additionally, it was proposed that leaving some gaps open may be a trade-off we're willing to make if the benefits brought by AI technologies in certain areas are significant enough. It's not necessary to adopt an all-or-nothing approach (Berber 2023). In this way, some of the responsibility gaps will persist alongside the technologies that give rise to them. This is another compelling reason to delve deeply into these gaps and their complexities.

Andrea Berber, Institute of Philosophy, Faculty of Philosophy, University of Belgrade
Jelena Mijić, Institute of Philosophy, Faculty of Philosophy, University of Belgrade

References:

Aristotle (2009). *The Nicomachean Ethics* (Oxford: Oxford University Press).
Baum, Kevin; Mantel, Susanne; Schmidt, Eva *et al.* (2022). "From Responsibility to Reason-Giving Explainable Artificial Intelligence" *Philosophy & Technology* 35, 12, https://doi.org/10.1007/s13347-022-00510-w.
Berber, Andrea (2023). "Automated decision-making and the problem of evil", *AI & Society*, https://doi.org/10.1007/s00146-023-01814-x
Berber, Andrea; Srećković, Sanja (2023). "When something goes wrong: who is responsible for errors in ML decision-making?", *AI & Society*,
https://doi.org/10.1007/s00146-023-01640-1.
Burrell, Jenna (2016). "How the machine 'thinks': Understanding opacity in machine learning algorithms", *Big Data & Society* 3(1), https://doi.org/10.1177/2053951715622512.
Danaher, John (2016). "Robots, law and the retribution gap", *Ethics and Information Technology* 18: 299-309.
Danaher, John (2022). "Tragic Choices and the Virtue of Techno-Responsibility Gaps", *Philosophy and Technology* 35, 26. https://doi.org/10.1007/s13347-022-00519-1.
de Jong, Roos (2020). "The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm", *Science and Engineering Ethics* 26: 727-735.
Elish, Madeleine Clare (2019). "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (pre-print)", *Engaging Science, Technology, and Society* 5: 40-60.
Feinberg, Joel (1962). "Problematic Responsibility in Law and Morals", *The Philosophical Review* 71 (3): 340-351.
Feinberg, Joel (1970). "Sua culpa", in: *Doing and Deserving; Essays in the Theory of Responsibility* (New Jersey: Princeton University Press): 187-223.
Fischer, John Martin; Ravizza, Mark (1998). *Responsibility and Control: A Theory of Moral Responsibility* (New York: Cambridge University Press).
Goodin, Robert E. (1995). *Utilitarianism as a Public Philosophy* (New York: Cambridge University Press).
Jeppsson, Sofia (2022). "Accountability, answerability and attributability. On different kinds of moral responsibility", in Dana Key Nelkin and Derk Pereboom (eds.), *The Oxford Handbook of Moral Responsibility* (New York: Oxford University Press): 73-88.
Lechterman, Theodore (2022). "The Concept of Accountability in AI Ethics and Governance", in Justin B. Bullock, and others (eds.), *The Oxford Handbook of AI Governance* (online edn, Oxford Academic),
https://doi.org/10.1093/oxfordhb/9780197579329.013.10.
List, Christian (2022). "Group Responsibility", in Dana Kay Nelkin and Derk Pereboom (eds.), *The Oxford Handbook of Moral Responsibility* (New York: Oxford University Press): 131-153.
Matthias, Andreas (2004). "The responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and Information Technology* 6: 175-183.
McKenna, Michael (2022). "Reasons-Responsiveness, Frankfurt Examples, and the Free Will Ability", in Dana Kay Nelkin and Derk Pereboom (eds.), *The Oxford Handbook of Moral Responsibility* (New York: Oxford University Press): 27-52.

McKinney Scott Mayer, Sieniek Marcin, Godbole Varun, *et al*. (2020). "International evaluation of an AI system for breast cancer screening", *Nature* 577: 89-94.
Oshana, Marina (1997). "Ascriptions of Responsibility", *American Philosophical Quarterly* 34 (1): 71-83.
Rudin, Cynthia (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature Machine Intelligence* 1: 206-215.
Santoni de Sio, Filippo; Mecacci, Giulio (2021). "Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them", *Philosophy and Technology* 34: 1057-1084.
Santoni de Sio, Fillippo; van den Hoven, Jaroen (2018). "Meaningful Human Control over Autonomous Systems: A Philosophical Account", *Frontiers in robotics and AI 5*, 15. https://doi.org/10.3389/frobt.2018.00015.
Smith, Helen; Birchley, Giles; Ives, Jonathan (2023). "Artificial intelligence in clinical decision-making: Rethinking personal moral responsibility", *Bioethics* 38: 78-86.
Smith, Helen (2021). "Clinical AI: opacity, accountability, responsibility and liability", *AI & Society* 36: 535-545.
Srećković, Sanja; Berber, Andrea; Filipović, Nenad (2022). "The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation", *Minds & Machines* 32: 159-183.
Shoemaker, David (2015). *Responsibility from the Margins* (Oxford: Oxford University Press).
Sparrow, Robert (2007). "Killer Robots", *Journal of Applied Philosophy* 24 (1): 62-77.
Thompson, Dennis F. (1980). "Moral Responsibility of Public Officials: The Problem of Many Hands", *American Political Science Review* 74 (4): 905-16.
Tu, Tao; Palepu, Anil; Schaekermann, Mike *et al*. (2024). "Towards Conversational Diagnostic AI", arXiv:2401.05654.
van de Poel, Ibo; Sand, Martin (2021). "Varieties of responsibility: two problems of responsible innovation", *Synthese* 198 (Suppl 19): 4769-4787.
van de Poel, Ibo (2015). "The Problem of Many Hands", in Ibo van de Poel, Lambèr Royakkers, Sjoerd D. Zwart (eds.), *Moral Responsibility and the Problem of Many Hands* (New York: Routledge): 50-92.
Vincent, Nicole (2009). "Responsibility: distinguishing virtue from capacity", *Polish Journal of Philosophy* 3 (1): 111-126.
United Nations, *Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011)*, March 8, 2021, S/2021/229, available at: https://undocs.org/Home/Mobile?FinalSymbol=S%2F2021%2F229&Language=E&DeviceType=Desktop&LangRequested=False (7 February 2024).
Watson, Gary (1996). "Two Faces of Responsibility", *Philosophical Topics* 24 (2): 227-248.
Williams, Garrath (2008). "Responsibility as a Virtue", *Ethical Theory and Moral Practice* 11 (4): 455-470.

Andrea Berber i Jelena Mijić

Razumevanje moralne odgovornosti u automatizovanom donošenju odluka: jazovi u odgovornosti i strategije za njihovo premošćavanje
(*Apstrakt*)

U ovom radu razmatramo upotrebu sistema zasnovanih na mašinskom učenju u procesima donošenja odluka i posledice po tradicionalno shvatanje moralne odgovornosti. Pažnja je posvećena pojavi jazova u odgovornosti i procenjivanju predloženih strategija za njihovo premošćavanje. Cilj rada je da pruži uvodni i sveobuhvatni pregled debata o moralnoj odgovornosti u automatizovanom donošenju odluka. Temeljnim razmatranjem ovih problema nastojimo da doprinesemo dubljem razumevanju implikacija integracije veštačke inteligencije u društvo.

KLJUČNE REČI: moralna odgovornost, jaz u odgovornosti, problem mnogih ruku, automatizovano donošenje odluka, objašnjiva AI, mašinsko učenje, veštačka inteligencija.